Dell EqualLogic Configuration Guide

Dell Storage Infrastructure and Solutions

- Configure unified block and file storage solutions based on EqualLogic PS Series arrays and the FS Series Family of NAS Appliances.
- Recommendations and best practices for iSCSI SAN and scale-out NAS network fabric design.
- Updated capacity guidelines, capabilities, limitations, and feature sets for the EqualLogic product family.

Version 13.2 June 2012

Abstract

This configuration guide provides technical guidance for designing and implementing Dell EqualLogic PS Series storage solutions in iSCSI SAN environments. The goal of this guide is to provide a single reference for product technical information and recommended SAN and NAS design methods:

- Details on the capacity and capabilities of different EqualLogic PS Series family and FS Series family of NAS Appliances
- An overview of the peer storage feature set provided by the EqualLogic storage controllers and the NAS feature set provided by the FS Series NAS Service
- iSCSI SAN design guidelines for optimizing redundancy, availability and performance of EqualLogic storage arrays.
- Host configuration guidelines and recommendations for Ethernet switch configuration

This document should be used strictly as a guide in planning an EqualLogic SAN solution. It should not be considered a statement of support for any specific configuration. Actual viability of any configuration will depend on the capabilities of the individual components (switches, network interface controllers, initiators, etc.) that make up the SAN infrastructure.

This configuration guide is intended for storage administrators, SAN designers, storage consultants, or anyone who is considering purchasing or has purchased an EqualLogic PS Series Array(s) for use in a production storage area network.

THIS WHITE PAPER IS FOR INFORMATIONAL PURPOSES ONLY, AND MAY CONTAIN TYPOGRAPHICAL ERRORS AND TECHNICAL INACCURACIES. THE CONTENT IS PROVIDED AS IS, WITHOUT EXPRESS OR IMPLIED WARRANTIES OF ANY KIND.

© 2012 Dell Inc. All rights reserved. Reproduction of this material in any manner whatsoever without the express written permission of Dell Inc. is strictly forbidden. For more information, contact Dell.

Dell, the *DELL* logo, and the *DELL* badge, *PowerConnect*™, *Dell™ EqualLogic*™, *PowerEdge*™ and *PowerVault*™ are trademarks of Dell Inc. *Broadcom*™ is a registered trademark of Broadcom Corporation. Cisco® is a registered trademarks of Cisco Systems, Inc. and/or its affiliates in the U.S. and other countries. *Intel*™ is a registered trademark of Intel Corporation in the U.S. and other countries. *Microsoft*™, *Windows*™, *Windows Server*™, and *Active Directory*™ are either trademarks or registered trademarks of Microsoft Corporation in the United States and/or other countries.

Revision History

Revision	Date	Changes
13.2	June 2012	Added PS Series Firmware Compatibility with EqualLogic Tools in Table 7.
13.1	March 2012	Updated for PS6110/4110. Updated Blade Integration in section 4.7. Updated the capacity tables from raw storage sizes to usable storage sizes. Added replication partner compatibility information.
12.4	November 2011	New sections and other changes throughout the document to include coverage of FS Series NAS Appliance (FS7500)
12.3	October 2011	Updates for PS4100/PS6100 family arrays: controller details; vertical port failover; SAN configuration guidelines; RAID level capacity tables
12.2.1	July 2011	Corrected Spanning Tree portfast guidance in Appendix D.
12.2	May 2011	Updated figures and content additions to Replication section
12.1	March 2011	Section 3.3.7: Replication Section 4.6: Integrating 1GbE and 10GbE SANs Section 2.1: Supported Configuration Limits (maintaining consistency with current firmware release notes) Appendix B and Appendix C: Content removed and now published as a separate **Validated Components List** document

All significant changes included in the most recent version of this document are highlighted using blue text.

Table of contents

1 Introduction 1

1.1 Audience 1

2 PS Series Storage Arrays 2

2.1 Array Models 2

2.2 PS Series Supported Configuration Limits 3

2.3 PS Series Storage Array Controllers 4

2.3.1 Controller Types in all models prior to PS4100/PS6100 4

2.3.2 Controller Redundancy in all models prior to PS4100/PS6100 5

2.3.3 Controller Failover behavior in all models prior to PS4100/PS6100 6

2.3.4 Controller Types in PS4100/PS6100 models 9

2.3.5 Controller Redundancy in PS4100/PS6100 controllers 9

2.3.6 Controller failover behavior: PS41x0/PS61x0 10

2.3.7 Vertical Port Failover behavior in PS4100/PS6100 Controllers 11

2.3.8 Vertical Port Failover behavior in PS4110/PS6110 Controllers 13

2.3.9 Controller Firmware 15

2.4 RAID Policies and Sizing Tables 17

2.4.1 RAID 5 Drive Layouts and Total Usable Storage 17

2.4.2 RAID 6 Drive Layouts and Total Usable Storage 18

2.4.3 RAID 10 Drive Layouts and Total Usable Storage 20

2.4.4 RAID 50 Drive Layouts and Total Usable Storage 21

3 PS Series Block Level Operations 22

3.1 Groups 22

3.2 Pools 22

3.3 Volumes 24

3.3.1 Volume Attributes 24

3.3.2 Volume Features 25

3.4 Thin Provisioning 25

3.5 Snapshots 26

3.6 Clones 27

3.7 Template volumes and thin clones 27

3.8 Replication 28

3.8.1 Replication Limits 28

3.8.2 Replication Paths 29
3.8.3 Replication Process 29
3.8.4 Fast failback 32
3.8.5 Sizing Replica Reserve and Delegated Space 32
3.8.6 Effect of TCP Window Size on Latency across WAN links 34
3.8.7 Replication Partner Compatibility 34
3.9 Clustering 35
4 EqualLogic SAN Design 36
4.1 Implementation of Standards 36
4.2 General Requirements and Recommendations 36
4.2.1 Quality of Service (QoS) 38
4.3 Ethernet Switches and Infrastructure 38
4.3.1 Connecting SAN Switches in a Layer 2 Network 39
4.3.2 Sizing Inter-Switch Connections 42
4.3.3 Comparing Inter-Switch Connection Types 42
4.4 Block I/O Host Configuration Guidelines 43
4.4.1 Design guidelines for host connectivity in a redundant SAN 44
4.4.2 Multi-Path I/O 45
4.5 EqualLogic iSCSI SAN Design 46
4.5.1 Redundant SAN Configuration 47
4.5.2 Partially Redundant SAN Configurations 48
4.5.3 Minimum Cabling Scenarios: PS4100 and PS6100 50
4.6 Integrating 1GbE and 10GbE SANs 52
4.6.1 Design considerations 53
4.6.2 Mixed SAN best practices 55
4.7 Blade Server Chassis Integration 55
4.7.1 Designing a SAN using Blade Chassis IO Modules with Directly-Attached Arrays 57
4.7.2 Designing a SAN using Blade Chassis IO Modules as Host Access to External Switches for Array Connection 65
4.7.3 M-Series Pass-Through Modules 70
5 FS Series NAS Appliance 71
5.1 FS Series Architecture 71
5.1.1 FS Series solution for file only storage 72
5.2 Dell FluidFS 73
5.3 FS Series Models 74

5.3.1 FS7500 Supported Configuration Limits 74
5.3.2 FS7500 System Components 76
5.3.3 FS7500 file system operation on controller failover 76
6 FS Series File Level Operations 78
6.1 NAS Service 78
6.2 NAS Reserve 78
6.2.1 Relationship between PS Series Groups, Pools and NAS Reserve 78
6.3 NAS File Systems 79
6.3.1 NAS File System security styles 80
6.4 File System Snapshots 81
7 FS Series NAS Configuration 83
7.1 FS7500 Connection Paths 83
Appendix A Network Ports and Protocols 86
A.1 Required Ports and Protocols 86
A.2 Optional Ports and Protocols 86
Appendix B Recommended Switches 88
Appendix C Supported iSCSI Initiators 89
Related Publications 90

1 Introduction

With the Dell™ EqualLogic™ PS Series of storage arrays Dell provides a storage solution that delivers the benefits of consolidated networked storage in a self-managing, iSCSI storage area network (SAN) that is affordable and easy to use, regardless of scale. By eliminating complex tasks and enabling fast and flexible storage provisioning, these solutions dramatically reduce the costs of storage acquisition and ongoing operations.

To be able to leverage the advanced features provided by an EqualLogic array, a robust, standards-compliant iSCSI storage area network (SAN) infrastructure must be created. While almost any industry standard, enterprise class Ethernet switch can be used to connect hosts to the EqualLogic-based virtual storage volumes, some switches, network interface controllers(NICs), iSCSI Host Bus Adapters (HBAs), and iSCSI initiators can directly impact the overall performance of the SAN storage solution. The *Dell EqualLogic Configuration Guide* is provided as an aid to help storage administrators determine how best to build an iSCSI infrastructure for use within an EqualLogic SAN solution. This document focuses on network configuration, host integration and other topics that help to ensure a smooth deployment with optimum PS Series SAN and FS Series NAS appliance performance.

Note: This version of the EqualLogic Configuration Guide (v13.1, March 2012) introduces coverage of the EqualLogic PS4110 and PS6110 second generation 10GbE arrays.
Unless otherwise stated, the recommendations involving specific 10GbE solutions apply to both the first generation PS6010 as well as the second generation PS4110 and PS6110.

Unless otherwise stated, recommendations in this document are applicable to both file and block environments.

This document should be used strictly as a guide in planning an EqualLogic storage solution. It should not be considered a statement of support for any specific configuration. Actual viability of any configuration will depend on the capabilities of the individual components (switches, network interface controllers, initiators, etc.) that make up the SAN infrastructure.

1.1 Audience

This configuration guide is intended for storage administrators, SAN/NAS system designers, storage consultants, or anyone who is considering purchasing or has purchased EqualLogic PS Series storage or FS Series Array appliances for use in a production storage area network. It is assumed that all readers have experience in designing and/or administering a shared storage solution. Also, there are some assumptions made in terms of familiarity with all current and possibly future Ethernet standards as defined by the Institute of Electrical and Electronic Engineers (IEEE) as well as all TCP/IP and iSCSI standards as defined by the Internet Engineering Task Force (IETF).

2 PS Series Storage Arrays

PS Series Storage SANs provide a peer storage architecture comprised of one or more independent arrays. Each array contains its own controllers, cache, storage, and interface ports. Grouped together they can create one or more single instance storage pools that are based on the IETF iSCSI standard. The capabilities of the PS Series solution are provided through a unique set of software and hardware capabilities that, when used cooperatively, can provide a full featured SAN solution. The following sections provide additional insight into specifications and functions of PS Series arrays.

2.1 Array Models

The array models available prior to introduction of the PS4100/PS6100 family are shown in Table 1 below.

Table 1 Array Models prior to PS4100/PS6100

Array Model	Drive Type	Number of Drives
PS50E – PS2400E	SATA	14, 16[(a)]
PS3000X, PS3x00XV	SAS	16
PS4000E	SATA	16
PS4000X, PS4000XV	SAS	16
PS5000E	SATA	16
PS5000X, PS5000XV	SAS	16
PS5500E	SATA	48
PS6000E	SATA	16
PS6000X, PS6000XV	SAS	16
PS6000S, PS6010S	SSD	16
PS6010E	SATA	16
PS6010X, PS6010XV	SAS	16
PS6000XVS, PS6010XVS	SAS / SSD	8 / 8
PS6500E	SATA	48
PS6510E	SATA	48
PS6510X	SAS	48

(a) PS2400E comprised of four drive enclosures with 14 drives each

Starting with the introduction of the PS4100 and PS6100 family of arrays, configurations using 2.5" and 3.5" disks are available.

Table 2 PS4100/PS6100 Array Models

Array Model	Drive Type	Number of Drives
PS4100E	3.5" SAS 7.2K RPM	12
PS4100X	2.5" SAS 10K RPM	24
PS4100XV	2.5" SAS 15K RPM	24
PS4100XV	3.5" SAS 15K RPM	12
PS6100E	3.5" SAS 7.2K RPM	24
PS6100X	2.5" SAS 10K RPM	24
PS6100XV	2.5" SAS 15K RPM	24
PS6100XV	3.5" SAS 15K RPM	24
PS6100S	SSD	12 or 24
PS6100XS	SSD + SAS 10K RPM	7 SSD + 17 SAS
PS4110E	3.5" SAS 7.2K RPM	12
PS4110X	2.5" SAS 10K RPM	24
PS4110XV	2.5" SAS 15K RPM	24
PS4110XV	3.5" SAS 15K RPM	12
PS6110E	3.5" SAS 7.2K RPM	24
PS6110X	2.5" SAS 10K RPM	24
PS6110XV	2.5" SAS 15K RPM	24
PS6110XV	3.5" SAS 15K RPM	24
PS6110S	SSD	12 or 24
PS6110XS	SSD + SAS 10K RPM	7 SSD + 17 SAS

2.2 PS Series Supported Configuration Limits

The Dell EqualLogic PS6xxx Series provides the full range of features and capabilities available with the EqualLogic PS Series storage products. The Dell EqualLogic PS4xxx Series provides a subset of features and capabilities, targeted at remote office and small to medium business storage deployments. The supported configuration limits for a PS Series group provided in Table 3.

Table 3 Supported Configuration Limits

Configuration	PS4000/PS4100 Groups Only[a]	All Other Groups[b]
Volumes and replica sets per group	512[h]	1024
Volume size[c]	15 TB	15 TB
Volumes enabled for replication (outbound)[d]	32	256
Snapshots and replicas per group	2048	10,000
Snapshots per volume	128	512
Replicas per volume	128	512
Schedules (snapshot or replication) per volume or volume collection	16	16
Persistent Reservation registrants per volume	96	96

Replication partners per group	16	16
Replication partners per volume	1	1
Members per group	2	16[a]
Members per pool	2	8
Pools per group	2	4
Volumes per collection	8	8
Collections per group (snapshot and replication)	100	100
Volume connections (each time an iSCSI initiator connects to a volume counts as a connection)[e,f]	512 per pool[g] 1024 per group with 2 pools	1024 per pool[h] 4096 per group with 4 pools
Access control records per volume and its snapshots	16	16
Simultaneous management sessions (any combination of GUI, telnet, or scripting sessions)	7	7
Thin provisioning[i] limits (minimum allocation)	10% of volume size	10% of volume size
Administrator accounts per group	100	100

(a) A group can contain a maximum of two (2) PS4000 and/or PS4100 arrays.
(b) Includes groups of mixed array types and all other group types except groups containing only PS4000 arrays. When a group contains a mix of PS4000 arrays and other array models, the higher limits prevail.
(c) Practical maximum volume size is operating system-specific. A PS Series group can create and present volumes up to 15 TB.
(d) Up to 16 simultaneous replication operations can take place.
(e) To avoid service outages or possible loss of connectivity during failovers, Dell recommends increasing initiator timeouts for groups with more than 512 connections. See the iSCSI Initiator and Operating Systems Considerations document for more information.
(f) Inbound replication connections count towards the total number of connections.
(g) With firmware version 5.1, maximum number of volumes per group and maximum number of connections per pool increased from 256 to 512
(h) Up to 1024 connections per pool are supported for all groups except groups containing only PS4000 arrays. This provides increased scalability in virtualized environments. These connections may be distributed across a maximum of 512 volumes.
(i) Volumes automatically created and managed by the NAS Service inside the NAS Reserve pool used by an FS Series appliance are fully allocated at creation and cannot be thin provisioned. Once those volumes are created (when the NAS Reserve is created they cannot be shrunk.

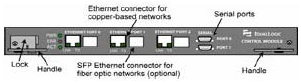
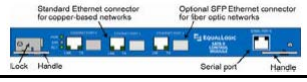
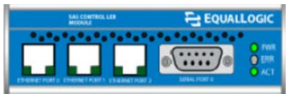
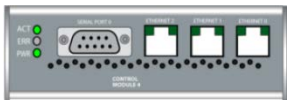
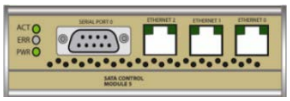
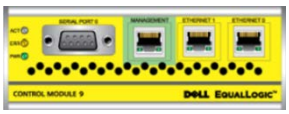
2.3 PS Series Storage Array Controllers

Since the EqualLogic PS Series was introduced, there have been several different array models released with new features, better performance and greater storage capacity. The storage array controllers were also improved to take advantage of advances in the underlying networking and storage technologies.

2.3.1 Controller Types in all models prior to PS4100/PS6100

Array controllers can be identified and differentiated by the controller "type" designation. Each controller type will have a different colored label to help quickly identify the controller type. Table 4 lists each Dell EqualLogic controller along with some characteristics about each.

Table 4 Array Controller Types – all models prior to PS4100/PS6100

Controller Type	Faceplate	Network Interfaces	Storage Type	Notes
Type 1		3 x 1GbaseT 3 x 1Gb SFP (combo)	SATA	• Original Controller Design • PS50 – PS2400 • 1GB Cache
Type 2		3 x 1GbaseT 3 x 1Gb SFP (combo)	SATA	• PS50 – PS2400 • 1GB Cache
Type 3 SAS **Type 3 SATA**		3 x 1GbaseT	SAS SATA	• PS3000 – PS5000 • 1GB Cache • Cannot mix Type 3 SAS with Type 3 SATA
Type 4		3 x 1GbaseT	SAS	• PS3000 – PS5000 • 1GB Cache • Cannot mix Type 3 SAS with Type 4 controller
Type 5		3 x 1GbaseT	SATA	• PS3000 – PS5000 • 1GB Cache • Cannot mix Type 3 SAS with Type 5 controller
Type 6		3 x 1GbaseT	SATA	• PS5500 only • 2GB Cache
Type 7		4 x 1GbaseT	SAS SATA SSD	• PS6000 – PS6500 • 2GB Cache • Cannot mix SAS and SATA drives in same array
Type 8		2 x1GbaseT 1 x 10/100Mb mgt	SAS SATA	• PS4000 only • 2GB Cache • Cannot mix SAS and SATA drives in same array
Type 9		2 x 1GbaseT 1 x 10/100Mb mgt	SAS SATA	• 2nd generation PS4000 • 2GB Cache • Cannot mix SAS and SATA drives in same array
Type 10		2 x 10GB SFP+ 1 x 10/100Mb mgt	SAS SATA SSD	• 10Gb Ethernet • PS6010 – PS6510 • 2GB Cache

2.3.2 Controller Redundancy in all models prior to PS4100/PS6100

Each array can be configured with either a single controller, or dual redundant controllers. The single controller configuration will provide the same level of I/O performance as a dual controller configuration. The dual controller configuration provides for redundancy. Redundant controllers will

prevent volume connections between hosts and SAN from being dropped in the event of an active controller failure.

The *Active Controller* is the controller which is processing all disk and network I/O operations for the array. A second controller in dual controller configurations will always be in a "passive" operating mode. In this mode, the secondary controller will exhibit the following characteristics:

- Each of its Ethernet ports are electrically inactive (active lights are off)
- The passive controller cache mirrors the cache of the active controller.

2.3.3 Controller Failover behavior in all models prior to PS4100/PS6100

To support redundant controller fail over, each Ethernet port on the active controller that is connected to the SAN must have its corresponding port on the passive controller also connected to the same SAN network. In the event of a controller failure, the passive controller will immediately activate and continue to process all data requests to the array. The following changes occur during fail over:

- The backup controller automatically enables each of the corresponding Ethernet ports that were enabled on the failed primary controller
- The IP addresses that were assigned to each of the failed controller Ethernet ports are reassigned to the corresponding ports on the second controller.

A link failure on one or more of the Ethernet ports on the active controller does not constitute a controller failure. For this reason, it is important to ensure that ports from each controller are connected to at least two different switches. This will prevent a switch failure from also disabling all paths between a host and its connected volumes.

It is also critical that port connections for both controllers are configured so that the *corresponding ports on each controller are connected to the SAN*. If port 0 on the active controller is the only port connected to the SAN, then you must use port 0 on the passive controller for the other connection to the SAN. This is illustrated in the partially connected scenario shown in Figure 1.

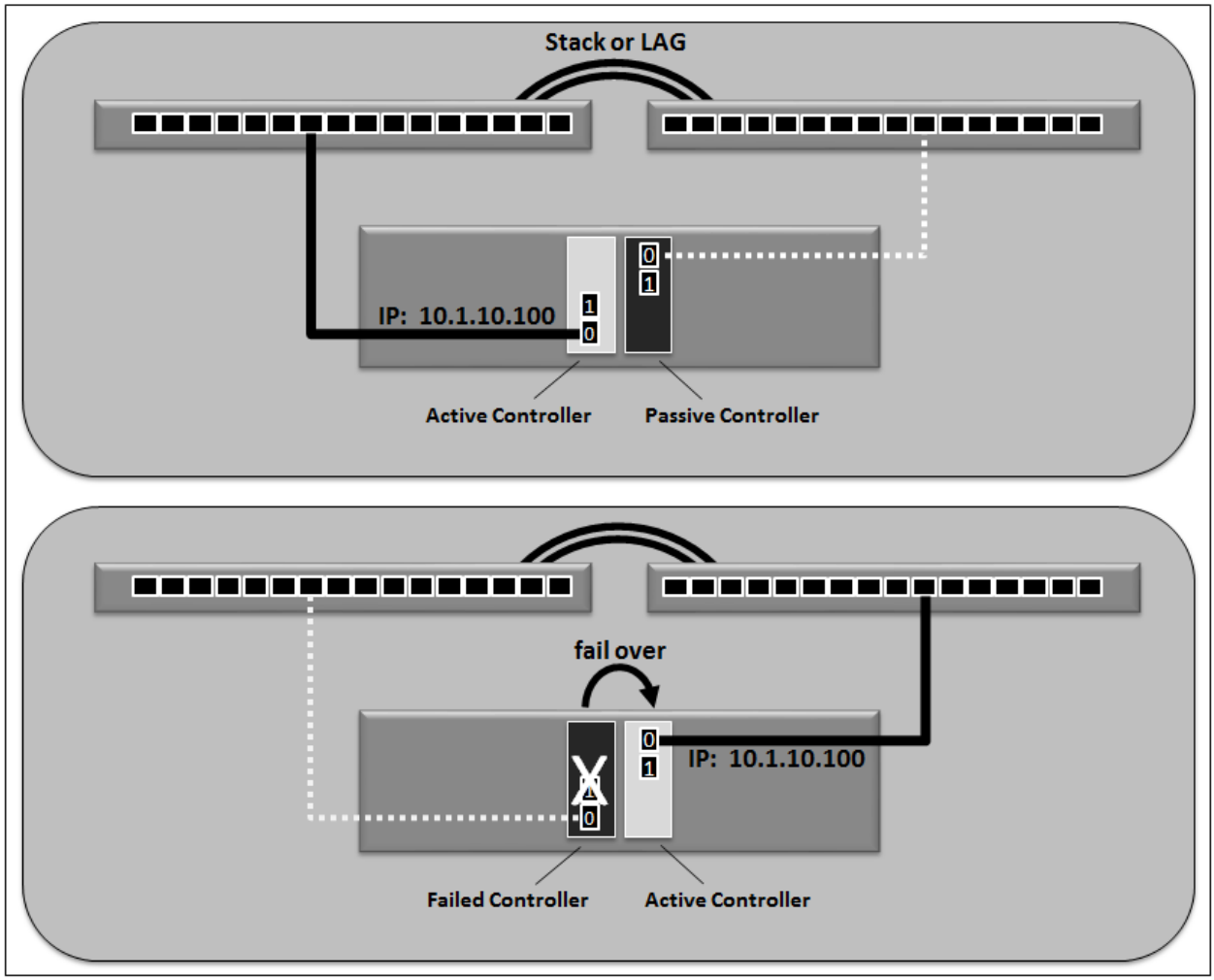


Figure 1 Partially Connected Controller Failover

Note how IP addresses are reassigned to the ports during the failover processes shown in Figure 1 and Figure 2.

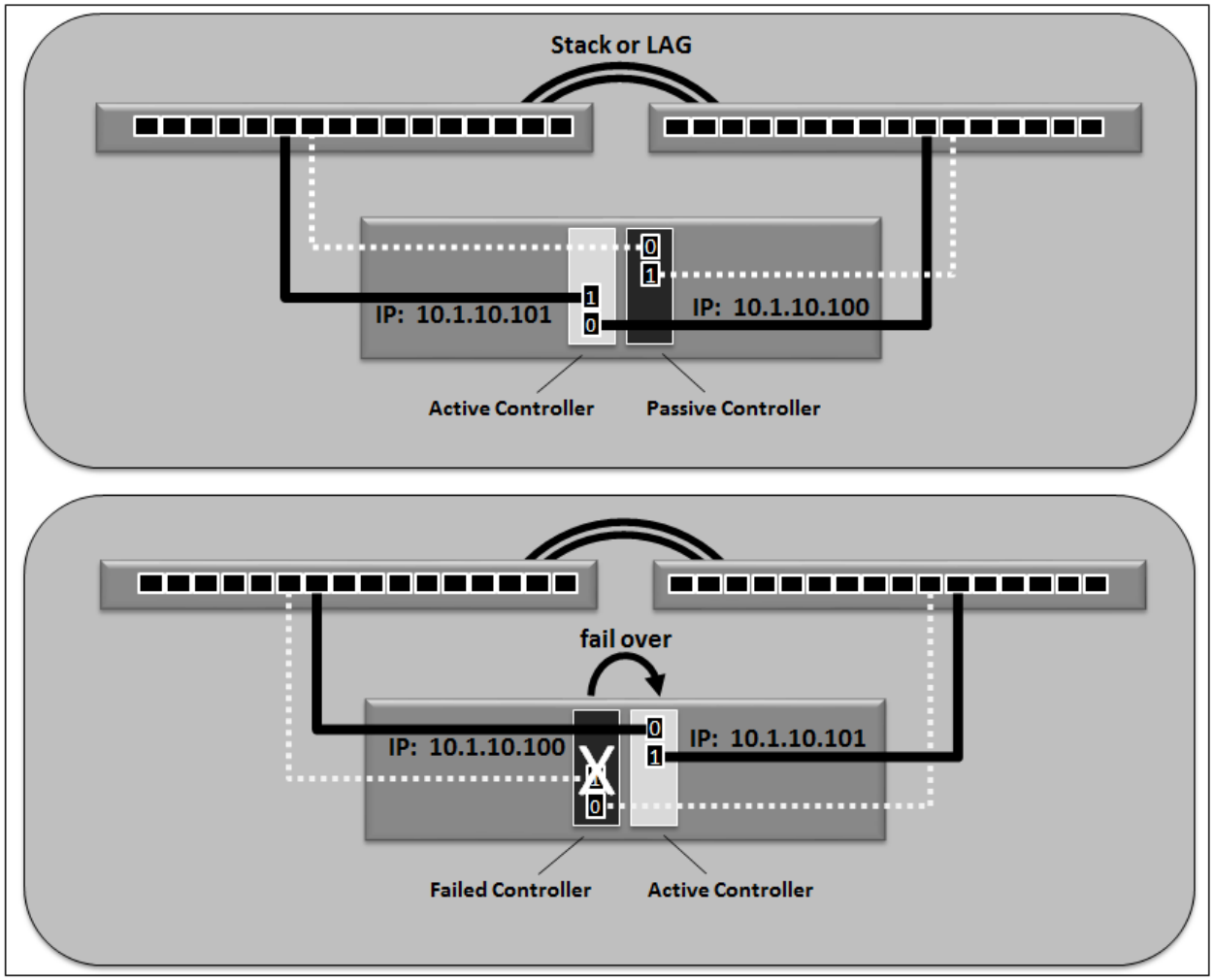


Figure 2 Fully Connected Controller Failover

2.3.4 Controller Types in PS4100/PS6100 models

The new controller types available in the PS4100 and PS6100 model arrays became available starting in August 2011. Table 5 lists each Dell EqualLogic controller along with some characteristics.

Table 5 PS4100/PS6100 controller types

Controller Type	Faceplate	Network Interfaces	Storage Type	Notes
Type 11		4 x 1GbaseT 1 x 10/100Mb mgt	SAS NL-SAS SSD	• PS6100 Only • 4GB Cache • Supports vertical port failover • Cache to Flash memory destaging
Type 12		2 x 1GbaseT 1 x 10/100Mb mgt	SAS NL-SAS	• PS4100 Only • 4GB Cache • Supports vertical port failover • Cache to Flash memory destaging
Type 14		1 x 10GbE SFP+ 1 x 10GbaseT 10/100Mb mgt	SAS NL-SAS SSD	• PS6110 Only • 4GB Cache • Support for vertical port failover • Cache to Flash memory destaging SFP+ OR 10GBaseT used at any given time.
Type 17		1 x 10GbE SFP+ 1 x 10GbaseT 1 x 10/100Mb mgt	SAS NL-SAS	• PS4110 Only • 4GB Cache • Supports vertical port failover • Cache to Flash memory destaging SFP+ OR 10GBaseT used at any given time.

2.3.5 Controller Redundancy in PS4100/PS6100 controllers

Each array can be configured with either a single controller, or dual redundant controllers. The single controller configuration will provide the same level of I/O performance as a dual controller configuration. The dual controller configuration provides for redundancy. Redundant controllers will prevent volume connections between hosts and SAN from being dropped in the event of an active controller failure.

The *Active Controller* is the controller which is processing all disk and network I/O operations for the array. A second controller in dual controller configurations will always be in a "passive" operating mode. In this mode, the secondary controller will exhibit the following characteristics:

- Each of its Ethernet ports are electrically inactive (active lights are off), unless a vertical port failover has occurred.
- The passive controller cache mirrors the cache of the active controller.

2.3.6 Controller failover behavior: PS41x0/PS61x0

In the event of a controller failure the following changes occur:

- The passive controller immediately activates and continues to process all data requests to the array.
- Vertical port pairing insures that IP addresses assigned to each of the failed controller Ethernet ports apply to the corresponding ports on the second controller.

As stated in Section 2.3.3 above, controller models prior to the PS4100/PS6100 required having cabled connections present on corresponding controller ports for controller failover to work correctly. The vertical port failover feature in PS41x0/PS61x0 controllers eliminates this requirement, enabling partially cabled configurations to support controller failover too. See Section 4.5.3, *Minimum Cabling Scenarios: PS4100 and PS6100*, for details.

It is important to ensure that ports from each controller are connected to at least two different switches. This will prevent a switch failure from also disabling all paths between a host and its connected volumes. We illustrate controller failover behavior for the PS4100 family controller in Figure 3. Controller failover behavior for the PS6100 (4 port) controller family is identical.

IMPORTANT: To prevent a switch failure from also disabling all paths between a host and its connected volumes, you should make sure that ports from each controller are connected to at least two different switches. **You should also split vertical port pair connections between two switches to ensure 100% bandwidth capability is maintained in the event of a vertical port failover event.** Both of these guidelines are illustrated in Figure 3.

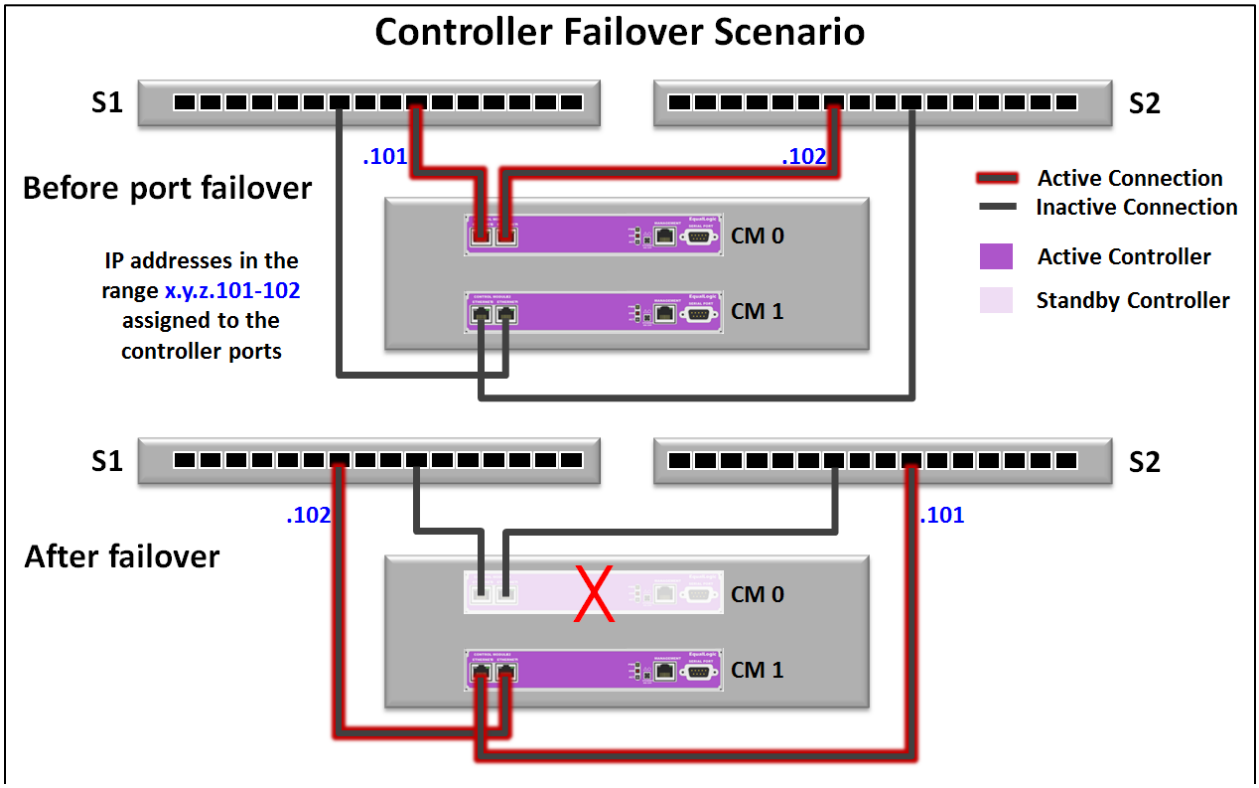


Figure 3 Controller Failover Process and Optimal Connection Paths

2.3.7 Vertical Port Failover behavior in PS4100/PS6100 Controllers

In PS Series controllers prior to PS4100/6100 families, a link failure or a switch failure was not recognized as a failure mode by the controller. Thus a failure of a link or an entire switch would reduce bandwidth available from the array. Referring to Figure 4 or Figure 5, assume that CM0 is the active controller. In vertical port failover, if CM0 senses a link drop on the local ETH0 port connection path, it will automatically begin using the ETH0 port on the backup controller (CM1). Vertical port failover is bi-directional. If CM1 is the active controller then vertical port failover will occur from CM1 ports to CM0 ports if necessary.

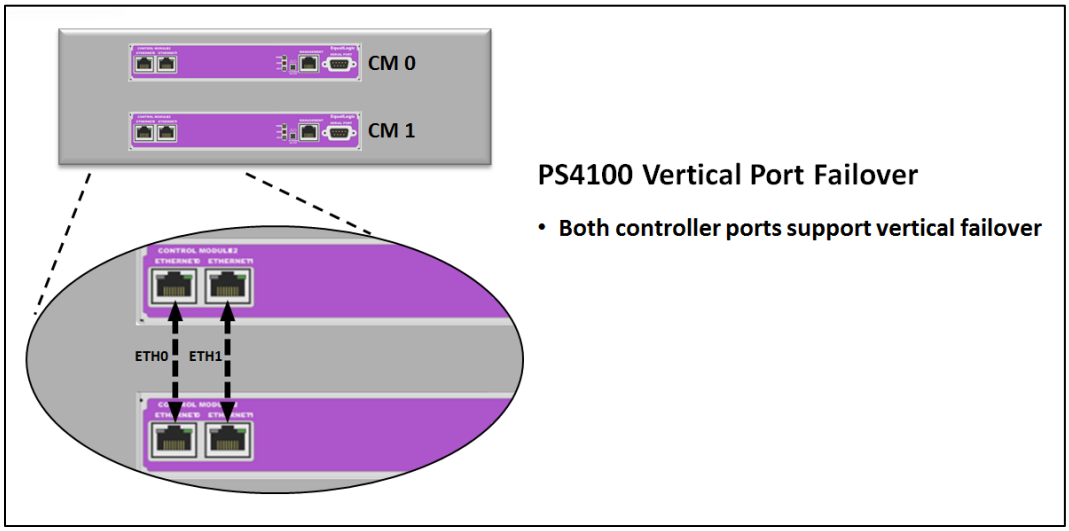


Figure 4 PS4100 Vertical Port Failover

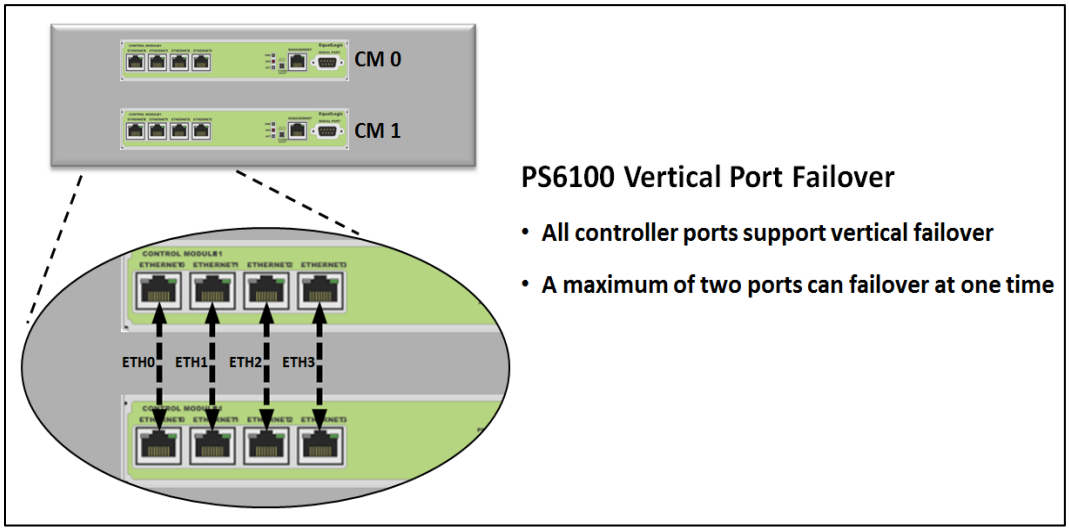


Figure 5 PS6100 Vertical Port Failover

With PS4100/PS6100 family controllers, vertical port failover can ensure continuous full bandwidth is available from the array even if you have a link or switch failure. This is accomplished by combining corresponding physical ports in each controller (vertical pairs) into a single logical port from the point of view of the active controller. In a fully redundant SAN configuration, you must configure the connections as shown in Figure 19 and Figure 20.

In a redundant switch SAN configuration, to optimize the system response in the event you have a vertical port failover **you must split the vertical port pair connections between both SAN switches**. The connection paths illustrated in Figure 6 and Figure 7 show how to alternate the port connection paths between the two controllers. Also note how IP addresses are assigned to vertical port pairs.

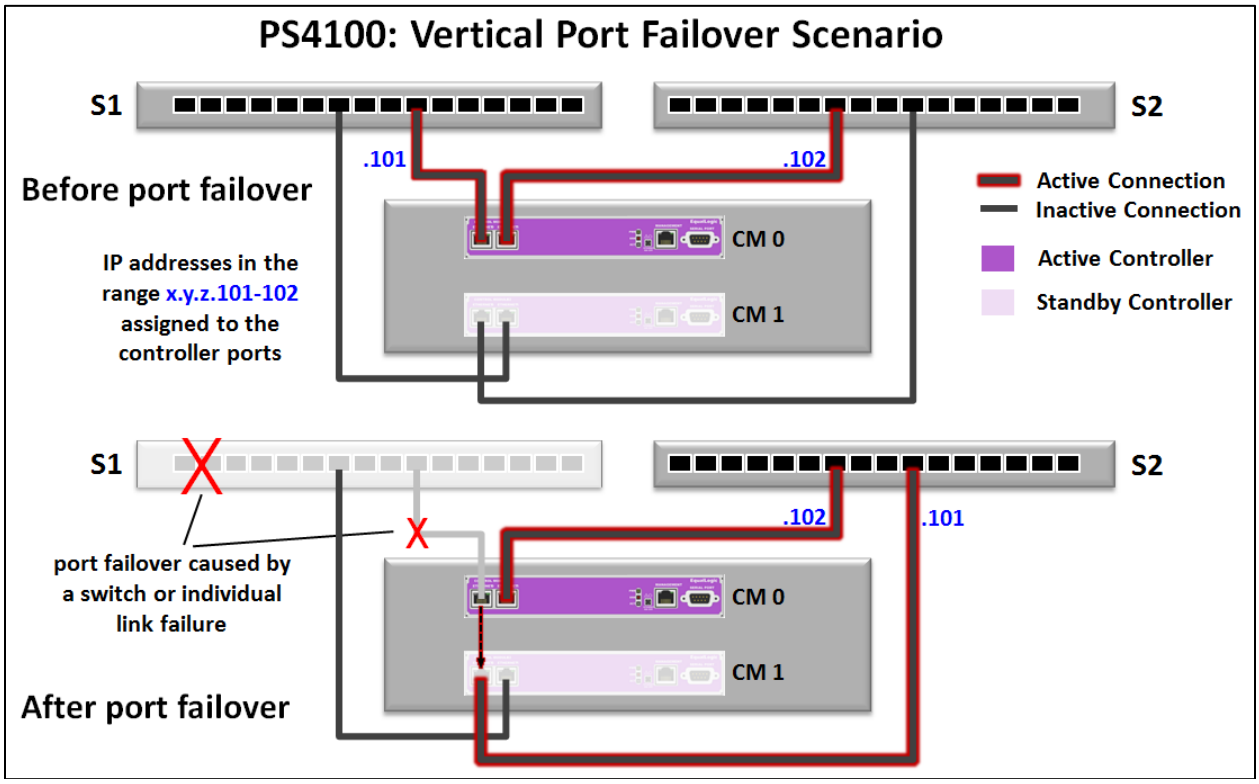


Figure 6 PS4100 Vertical Port Failover and Optimal Connection Paths

IMPORTANT: By alternating switch connection paths between ports in a vertical port pair, port failover allows the array to maintain 100% bandwidth capability in the event of a switch failure.

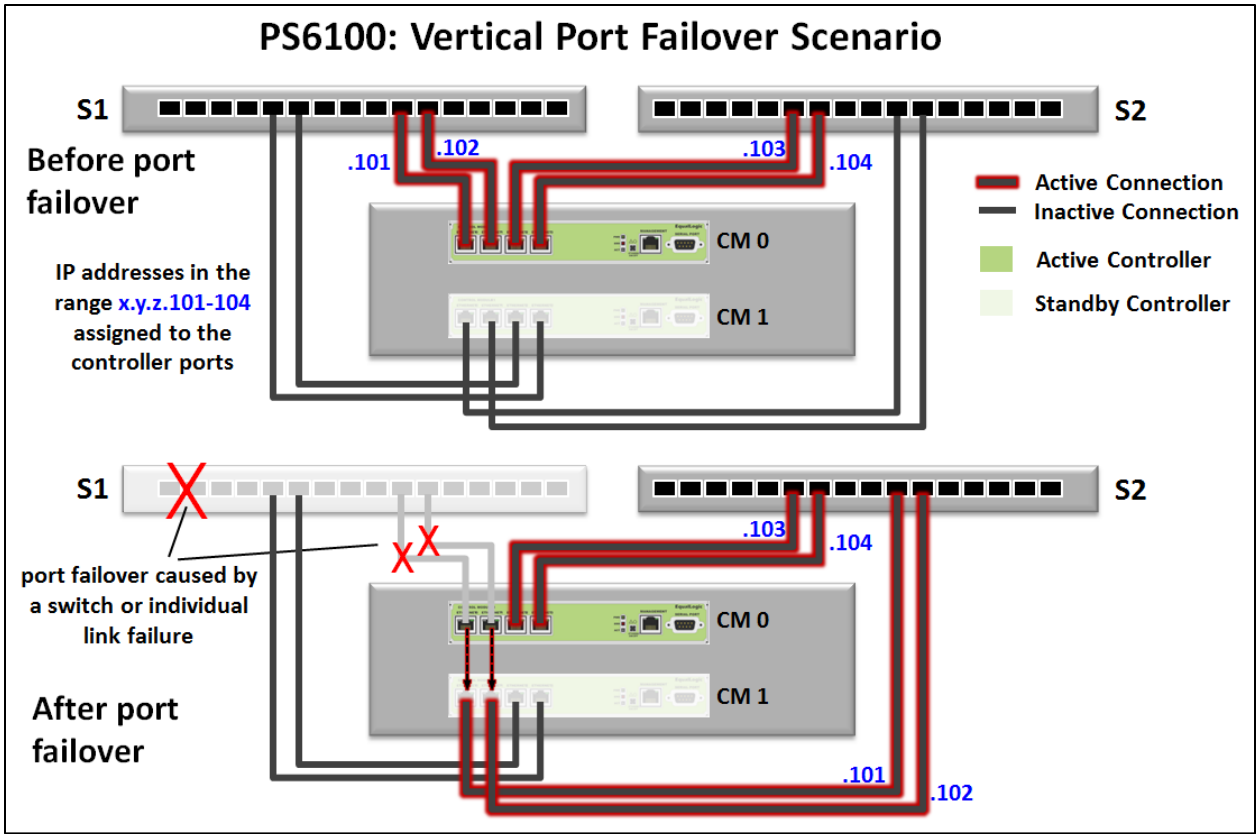


Figure 7 PS6100 Vertical Port Failover Process and Optimal Connection Paths

2.3.8 Vertical Port Failover behavior in PS4110/PS6110 Controllers

In PS Series controllers prior to PS4110/6110 families, a link failure or a switch failure was not recognized as a failure mode by the controller. This caused a failure of a link or an entire switch to reduce bandwidth available from the array. Referring to Figure 4 or Figure 5, assume that CM0 is the active controller. In vertical port failover, if CM0 senses a link drop on the local ETH0 port connection path, it will automatically begin using the ETH0 port on the backup controller (CM1). Vertical port failover is bi-directional. If CM1 is the active controller then vertical port failover will occur from CM1 ports to CM0 ports if necessary.

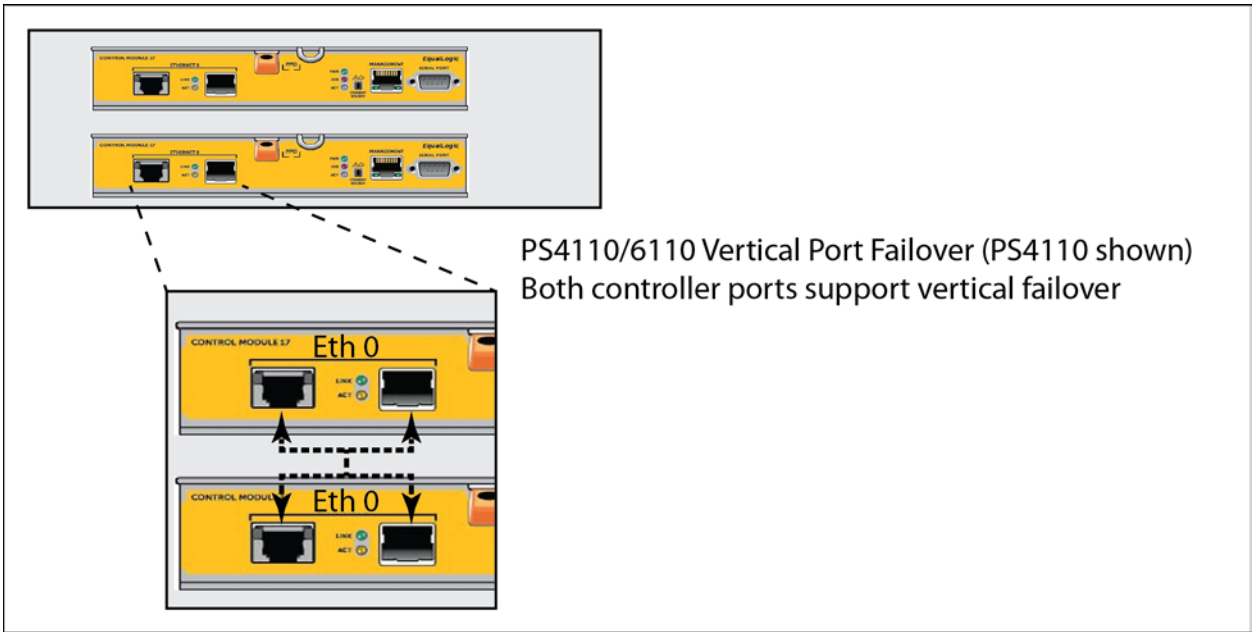


Figure 8 4110/6110 Vertical Port Failover

With the PS4110/PS6110 family of controllers, vertical port failover can ensure continuous full bandwidth is available from the array even if you have a link or switch failure. This is accomplished by combining 10GbE "eth0" ports in each controller into a single logical port from the point of view of the active controller. In a fully redundant SAN configuration, you must configure the connection as shown in Figure 9.

In a redundant switch SAN configuration, to optimize the system response in the event you have a vertical port failover **you must connect either the SFP+ cable or the 10GbaseT cable from each controller to a different switch in the SAN network**. While it is a supported configuration, it is not recommended to connect both the SFP+ and 10GbaseT ports on each controller at the same time. In this scenario, the SFP+ port will always be the preferred active port and this preference cannot be changed by the user. This preference is not guaranteed in future array firmware releases.

The connection paths illustrated in Figure 9 show how to connect the port connection paths between the two controllers. Also note how IP addresses are assigned to the vertical port pair.

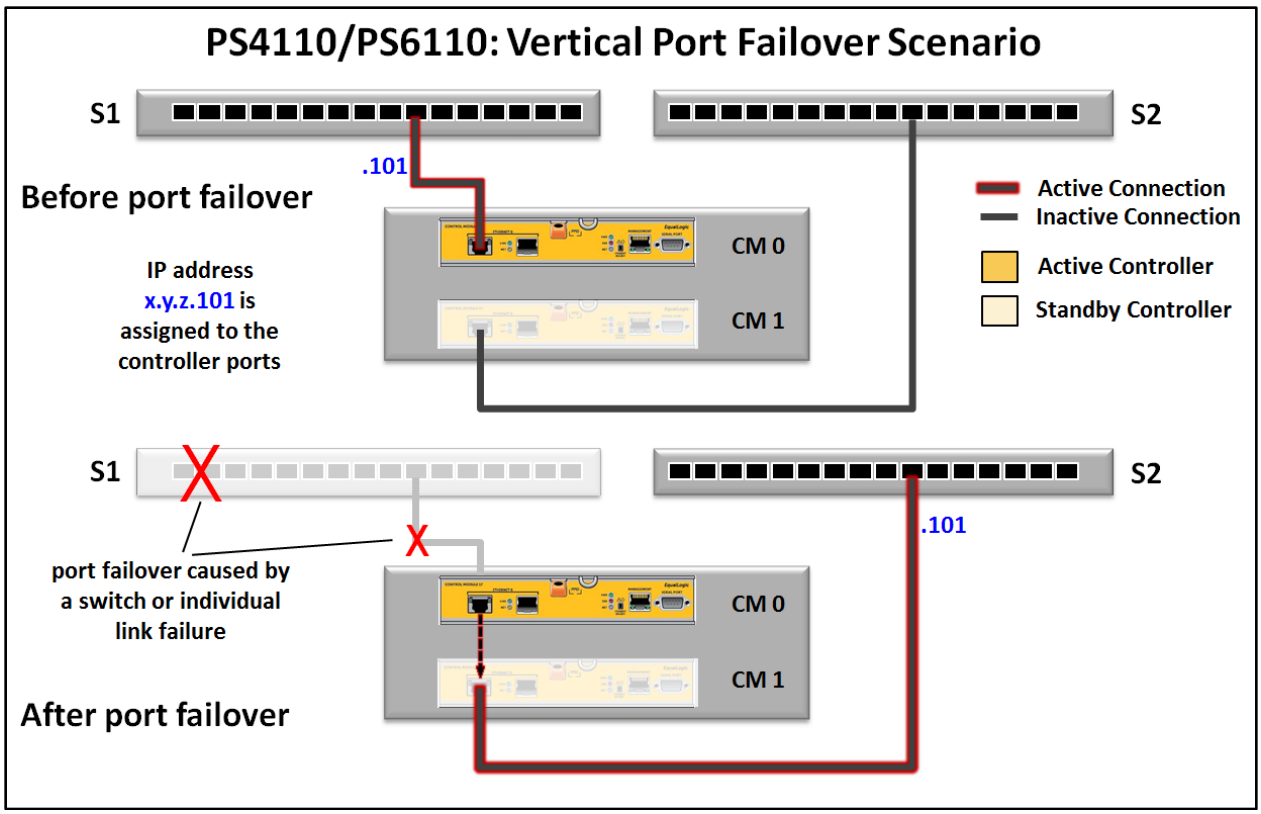


Figure 9 4110/6110 Vertical Port Failover Scenario

2.3.9 Controller Firmware

Each EqualLogic PS Series array runs a core operating system in firmware that provides all of the PS Series features and functionality. The firmware version is defined using a version number and will be updated from time to time as new features are added or for general maintenance improvements.

The firmware version number takes the following form: "*X.Y.Z*":

- "*X*" is used to identify the "major" release number. This changes when there are very significant changes in the features and functionality.
- "*Y*" is the "minor" release number. This changes as new minor features or improvements in existing features are made for a given major release.
- "*Z*" denotes maintenance releases. This changes whenever a maintenance release level is released to resolve any identified technical or usability issues.

2.3.9.1 Firmware Upgrade

Before beginning a firmware upgrade process, you must review the Release Notes for the firmware update revision. The release notes will be included with the firmware update kit. All firmware update kits are available at http://www.equallogic.com/support/ (Support ID required for login access). The release notes contain the latest information and guidelines for:

- Completing the upgrade with minimal potential for service disruption
- Operating system and iSCSI initiator configuration settings that will minimize potential for I/O disruption during the upgrade process
- Latest updates to the supported firmware upgrade paths

In addition to the Release Notes, the process for updating controller firmware is described in detail in the following document (Support ID required for login access):

- *PS Series Storage Arrays: Updating Storage Array Firmware*, available at: https://www.equallogic.com/support/download_file.aspx?id=594
- Supported firmware upgrade paths (up to version 5.0.x) are shown in Table 6 below.
- If you are starting with v4.2.* or v4.3.* then you can update straight to v5.0.4. If the array is already running v5.0 or v5.0.1 then you must first update to v5.0.2 before updating to v5.0.4.

Table 6 Controller Firmware Upgrade Paths

Starting Firmware Version	Ending Firmware Version
5.2.x	Later 5.2.x releases
5.1.x	Later 5.1.x releases 5.2.x
5.0.x (excluding 5.0.0 and 5.0.1)	Later 5.0.x releases 5.1.x, 5.2.x
5.0.0, 5.0.1	5.0.2
4.3.x	Later 4.3.x releases 5.0.x (excluding 5.0.0 & 5.0.1)
4.2.x	Later 4.2.x releases 4.3.x 5.0.x (excluding 5.0.0 & 5.0.1)
4.1.x	Later 4.1.x releases 4.2.x, 4.3.x
4.0.x	Later 4.0.x releases 4.1.x
3.3.x	Later 3.3.x releases 4.0.x
3.2.x	Later 3.2.x releases 3.3.x 4.0.x
3.1.x	Later 3.1.x releases 3.2.x, 3.3.x
3.0.x	Later 3.0.x releases 3.1.x

2.3.9.2 PS Series Firmware Compatibility with EqualLogic Tools

The following table provides a quick reference of EqualLogic product version compatibility for the recent major firmware releases.

Table 7 PS Series Firmware Compatibility

	Firmware V5.2.x	Firmware V5.1.x	Firmware V5.0.x	Firmware V4.3.x
SAN HeadQuarters	2.2.x 2.1.x	2.2.x 2.1.x	2.2.x 2.1.x	2.2.x 2.1.x
Host Integration Tools for Microsoft	4.0.x 3.5.x	4.0.x 3.5.x	4.0.x 3.5.x 3.4.x 3.3.x	4.0.x 3.5.x 3.4.x 3.3.x
Host Integration Tools for VMware	3.1.x 3.0.x	3.1.x 3.0.x 2.0.x	3.1.x 3.0.x 2.0.x	2.0.x
EqualLogic Storage Replication Adapter for VMware Site Recovery Manager	2.1.x 2.0.x 1.0.x	2.1.x 2.0.x 1.0.x	2.1.x 2.0.x 1.0.x	1.0.x
EqualLogic Multipathing Extension Module for VMware vSphere	1.1.x 1.0.x	1.1.x 1.0.x	1.1.x 1.0.x	1.1.x 1.0.x
Host Integration Tools for Linux	1.1.x 1.0.x	1.1.x 1.0.x	1.1.x 1.0.x	1.0.x
Manual Transfer Utility	1.2.1	1.2.1	1.2.1 1.1.2 (Windows only) 1.0.2 (Linux only)	1.2.1 1.1.2 (Windows only) 1.0.2 (Linux only)

2.4 RAID Policies and Sizing Tables

Each array in an EqualLogic array group is configured with a single RAID policy. Arrays (or group members) within the same storage pool that have the same RAID policy will cooperatively work to host volumes by distributing those volumes over multiple arrays. Two things that are defined by the RAID policy are:

- RAID level
- hot-spare configuration

Each array implements a default RAID policy that includes a hot-spare. To configure a RAID policy that does not include a hot-spare, you must manually initialize the array using the Command Line Interface.

Note: Configuring a RAID policy that does not include a hot-spare will increase the risk of data loss in the event of multiple drive failures.

2.4.1 RAID 5 Drive Layouts and Total Usable Storage

RAID 5 (striped disks with distributed parity) will combine N disks in an arrangement where each stripe consists of N–1 disks that contain data blocks plus 1 disk that contains a parity block. For each stripe, the parity block will be placed on a different disk ensuring that the parity blocks are not located on a single disk in the RAID set. RAID 5 implementations can tolerate a single drive failure without data loss.

Table 8 shows the drive layouts that are enforced when using a RAID 5 policy based on the number of drives in each array/hot spare configuration, and the total usable storage available for each model.

Table 8 RAID 5 Drive Layouts and Total Storage Available with Hot Spares (in GB)

Disk Drives	Hot Spare	No Hot Spare
6	5 Data/Parity + 1 Hot Spare	RAID Policy for RAID 5 without a hot spare is not supported
7	6 Data/Parity + 1 Hot-spare	
8	7 Data/Parity + 1 Hot-spare	
12[a]	11 Data/Parity + 1 Hot-spare	
14	13 Data/Parity + 1 Hot-spare	
16	15 Data/Parity + 1 Hot-spare	
24[a][b]	23 Data/Parity+ 1 Hot-spare	
48[b]	46 Data/Parity + 2 Hot-spares	

(a) PS4100/PS6100 models only
(b) 24 and 48 drive arrays implement multiple RAID 5 sets within a chassis

Total Usable Storage Available: All Models Prior to PS4100/PS6100

Disk Drive Qty / Size	50[a]	100[a]	74	146	250	300	400	450	500	600	750	1000	2000	3000
7[b]	-	-	-	-	1193	-	-	-	-	-	-	-	-	-
8[c]	-	-	-	-	1432	-	-	-	-	-	-	-	-	-
14[d]	-	-	851	-	2868	-	4597	-	5749	-	-	-	-	-
16	667	1336	-	1962	3348	4020	5364	6037	6709	8053	10070	13430	-	41297
48	-	-	-	-	-	-	-	-	20887	25067	-	41786	80079	123903

Total Usable Storage Available: PS4100/PS6100

Disk Drive Qty / Size	146	200[e]	300	400[e]	500	600	900	1000	2000	3000
6	-	-	-	-	1913	2286	-	3826	7817	11819
12	1396	1979	2869	3969	4782	5750	8631	9587	19527	29499
24	2946	4167	6036	8345	10071	12083	18097	20110	41000	61957

(a) Solid State Disk
(b) Legacy PS50E
(c) PS4000E and PS6000E only
(d) Legacy PS70E,PS100E, PS100X, PS200E and PS300E
(e) PS6100S (SSD) only

2.4.2 RAID 6 Drive Layouts and Total Usable Storage

RAID6 (striped set with dual distributed parity) combines N disks in an arrangement where each stripe consists of N-2 disks containing data blocks and 2 disks containing parity blocks. Each parity block generates parity using a different view of the data blocks depending on the RAID 6 implementation. RAID 6 can tolerate up to 2 drive failures per RAID set at the same time without data loss. RAID 6 is not recommended for workloads consisting mainly of random writes. Using a RAID 6 policy, Table 9

shows the drive layouts that are enforced based on the number of drives in each array/hot spare configuration, and the total usable storage available for each model.

Table 9 RAID 6 Drive Layouts and Total Storage Available with Hot Spares (in GB)

Disk Drives	Hot Spare	No Hot Spare
6	5 Data/Parity + 1 Hot-spare	6 Data/Parity
7	6 Data/Parity + 1 Hot-spare	7 Data/Parity
8	7 Data/Parity + 1 Hot-spare	8 Data/Parity
12[f]	11 Data/Parity + 1 Hot-spare	12 Data/Parity
14	13 Data/ Parity + 1 Hot-spare	14 Data/Parity
16	15 Data/ Parity + 1 Hot-spare	16 Data/Parity
8 + 8[c]	15 Data/Parity[d] + 1 Hot-spare	16 Data/Parity
24[a][f]	23 Data/Parity + 1 Hot-spare	24 Data/Parity
7 + 17[e]	23 Data/Parity + 1 Hot-spare	24 Data/Parity
48[a]	47 Data/Parity + 1 Hot-spare	N/A[b]

(a) 24 and 48 drive arrays implement multiple RAID 6 sets within a single chassis.
(b) 48 drive arrays cannot implement a no hot-spare RAID policy.
(c) PS60x0 XVS with 8x 100GB SSD and 8x 450GB SAS.
(d) One 450GB hot spare, 2x 450 GB parity and 2x 100GB SSD parity. In the event of a drive failure (SSD or SAS), the RAID set is reconstructed using the 450GB SAS hot spare.
(e) One 600GB hot spare, 2x 600 GB parity and 2x 400 GB SSD parity. In the event of a drive failure (SSD or SAS), the RAID set is reconstructed using the 600 GB SAS hot spare.
(f) PS4100/PS6100 models only

Total Usable Storage Available: All Models Prior to PS4100/PS6100

Disk Drive Qty / Size	50[a]	100[a]	74	146	250	300	400	450	500	600	750	1000	2000	3000
7[b]	-	-	-	-	954	-	-	-	-	-	-	-	-	-
8[c]	-	-	-	-	1193	-	-	-	-	-	-	-	-	-
14[d]	-	-	709	-	2388	-	3828	-	4789	-	-	-	-	-
16	619	1240	-	1821	3108	3732	4980	5605	6229	7477	9350	12470	-	38350
8 + 8[e]	-	-	-	-	-	-	-	2726	-	-	-	-	-	-
48	-	-	-	-	-	-	-	-	19394	23275	-	38800	74367	115052

Total Usable Storage Available: PS4100/PS6100

Disk Drive Qty / Size	146	200	300	400	500	600	900	1000	2000	3000
6	-	-	-	-	1429	1715	-	2869	5860	8862
12	1264	1781	2583	3573	4310	5167	7762	8631	17570	26542
24	2660	3771[f]	5453	7553[f]	9103	10929	16404	18196	37097	56053
7 + 17						10005[g]				

(a) Solid State Disk
(b) Legacy PS50E
(c) PS4000E and PS6000E only
(d) Legacy PS70E, PS100E, PS100X, PS200E, PS300E and PS400E
(e) PS60x0 XVS with 8 x 100GB SSD and 8 x 450GB SAS
(f) PS6100S (SSD) only
(g) PS6100XS only

2.4.3 RAID 10 Drive Layouts and Total Usable Storage

Using a RAID 10 policy, Table 10 shows the drive layouts that are enforced based on the number of drives in each array/hot spare configuration, and the total usable storage available for each model.

RAID 10 (mirrored sets in a striped set) combines two high performance RAID types: RAID 0 and RAID 1. A RAID 10 is created by first building a series of two disk RAID 1 mirrored sets, and then distributing data over those mirrors. RAID 10 implementations can tolerate one drive failure per mirror pair.

Table 10 RAID 10 Drive Layouts and Total Storage Available with Hot Spares (in GB)

Disk Drives	Hot Spare	No Hot Spare
6	4 Data (2 mirrors) + 2 Hot-spares	6 Data (3 mirrors) + 0 Hot-spare
7	6 Data (3 mirrors) + 1 Hot-spare	N/A[a]
8	6 Data (3 mirrors) + 2 Hot-spares	8 data (4 mirrors)
12[c]	10 Data (5 mirrors) + 2 Hot-spares	12 data (7 mirrors)
14	12 Data (6 mirrors) + 2 Hot-spares	14 data (7 mirrors)
16	14 Data (7 mirrors) + 2 Hot-spares	16 data (8 mirrors)
24[c]	22 Data (11 mirrors) + 2 Hot-spares	24 data (12 mirrors)
48	46 Data (23 mirrors) + 2 Hot-spares	N/A[b]

(a) Not supported. An even number of drives is required by RAID 10
(b) 48 drive arrays cannot implement a no hot-spare policy.
(c) PS4100/PS6100 models only

Total Usable Storage Available: All Models Prior to PS4100/PS6100

Disk Drive Qty / Size	50[a]	100[a]	74	146	250	300	400	450	500	600	750	1000	2000	3000
7[b]	-	-	-	-	715	-	-	-	-	-	-	-	-	-
8[c]	-	-	-	-	715	-	-	-	-	-	-	-	-	-
14[d]	-	-	425	-	1432	-	2292	-	2868	-	-	-	-	-
16	332	667	-	980	1670	2004	2676	3012	3348	4020	5029	6709	13688	20648
48	-	-	-	-	-	-	-	-	11432	13721	-	22877	43851	67850

Total Usable Storage Available: PS4100/PS6100

Disk Drive Qty / Size	146	200[e]	300	400[e]	500	600	900	1000	2000	3000
6	-	-	-	-	953	1143	-	1913	3903	5904
12	698	988	1429	1979	2385	2869	4310	4782	9774	14777
24	1539	2177	3155	4365	5266	6322	9488	10544	21473	32446

(a) Solid State Disk
(b) Legacy PS50E
(c) PS4000E and PS6000E only
(d) Legacy PS70E,PS100E, PS100X, PS200E and PS300E
(e) PS6100S (SSD) only

2.4.4 RAID 50 Drive Layouts and Total Usable Storage

Table 11 shows the drive layouts that are enforced when using a RAID 50 policy based on the number of drives in each array/hot spare configuration and the total usable storage available for each model.

RAID 50 (RAID 5 sets in a striped set) is created by first creating two or more RAID 5 sets and then striping data over those RAID5 sets. RAID 50 implementations can tolerate a single drive failure per RAID5 set.

Table 11 RAID 50 Drive Layouts and Total Storage Available with Hot Spares (in GB)

Disk Drives	Hot Spare	No Hot Spare
6	5 Data/Parity + 1 Hot-spare	N/A[a]
7	6 Data/Parity + 1 Hot-spare	N/A[a]
8	6 Data/Parity + 2 Hot-spares	8 data/parity
12[d]	10 Data/Parity + 2 Hot-spares	12 data/parity
14	12 Data/Parity + 2 Hot-spares	14 data/parity
16	14 Data/Parity + 2 Hot-spares	16 data/parity
24[d]	22 Data/Parity + 2 Hot-spares	24 data/parity
48[b]	46 Data/Parity + 2 Hot-spares	N/A[c]

(a) RAID 50 requires an even number of disk drives. A 7 drive configuration without hot-spare would result in odd number of disk drives.
(b) 48 drive arrays implement stripes across multiple RAID 5 sets within a single chassis.
(c) 48 drive arrays cannot implement a no hot-spare policy.
(d) PS4100/PS6100 models only

Total Usable Storage Available: All Models Prior to PS4100/PS6100

Disk Drive Qty / Size	50[a]	100[a]	74	146	250	300	400	450	500	600	750	1000	2000	3000
7[b]	-	-	-	-	954	-	-	-	-	-	-	-	-	-
8[c]	-	-	-	-	954	-	-	-	-	-	-	-	-	-
14[d]	-	-	709	-	2388	-	3828	-	4789	-	-	-	-	-
16	572	1145	-	1681	2868	3444	4596	5173	5749	6901	8629	11510	-	35404
48	-	-	-	-	-	-	-	-	19394	23275	-	38800	74367	115052

Total Usable Storage Available: PS4100/PS6100

Disk Drive Qty / Size	146	200[e]	300	400[e]	500	600	900	1000	2000	3000
6	-	-	-	-	1913	2286	-	3826	7817	11819
12	1121	-	2286	3166	3826	4595	6904	7663	15646	23595
24	2628	3573	5167	7157	8631	10357	15547	17273	35140	53106

(a) Solid State Disk
(b) Legacy PS50E
(c) PS4000E and PS6000E only
(d) Legacy PS70E,PS100E, PS100X, PS200E and PS300E
(e) PS6100S (SSD) only

3 PS Series Block Level Operations

3.1 Groups

A PS Series SAN Group is a Storage Area Network (SAN) comprised of one or more PS Series arrays connected to an IP network. Each array in a group is called a group *member*. Each member is assigned to a storage pool. There can be up to 4 pools within the group.

A group can consist of up to 16 arrays of any family or model as long as all arrays in the group are running firmware with the same major and minor release number. For example, it is supported to have different arrays in the same group running different maintenance release levels, as long as their major and minor revision levels match. An example of this would be running version 4.0.1 and 4.0.3 on different arrays in the same group. An example of an unsupported configuration would be different arrays in the same group running 4.0.1 and 4.1.0 (the minor versions differ). The only exception to this rule is for short term time periods when array firmware versions may be out of sync while upgrading the firmware on each array within a group. Features available for use within the group are determined by the lowest version of FW running within the group.

Note: We recommended that all arrays run the same version of PS Series firmware at all times, except during a firmware upgrade process. (See Section 2.3.9 on page 15 for more information.)

3.2 Pools

A pool is a container that each member is assigned after being added to the group. A pool can have between 1 to 8 members. There is always at least one pool in any group and it is called the "default pool" unless the name is changed. Regardless of the name of this pool, it is always considered the default storage pool. All newly added members are automatically assigned to the default pool. The default pool cannot be deleted.

Pools can be used as part of an overall tiered storage strategy for the SAN. Tiered storage is storage that is differentiated and optimized for the type of applications and data being hosted. Instead of optimizing all of the storage in the group for maximum storage utilization, a tiered storage system allows for the administrator to optimize different arrays for a variety of requirements such as application performance requirements or cost efficiency.

Pools are the containers that provide scope for all automated storage optimization features of the group. Pools with multiple arrays can implement different RAID policies within the pool. The EqualLogic automated optimization features will automatically move a volume from an array or set of arrays with one RAID policy to another array or set of arrays with a different RAID policy if it is determined that the application needs more (or less) performance.

The following rules apply to pools:

- Each member can be assigned to only one pool.
- Each pool can host up to 8 members.
- Each group will have at least 1 pool – the default pool.
- Each group can have up to 4 pools.

- If all members in the pool are running PS Series firmware v5.0 or later then you can mix PS5500E, PS6500E/X and PS6510E/X models together with other array models in the same pool.
- If you are running PS Series firmware version prior to v5.0 then PS5500E, PS6500E/X and PS6510E/X models must reside in a separate pool from other array types.

Figure 3 shows a PS Series group with the maximum of 4 pools. Note the use of Pool 3 for containing PS5500/PS6500 series arrays only. Also note that Pool 3 contains arrays that implement both 1GbE and 10GbE controllers.

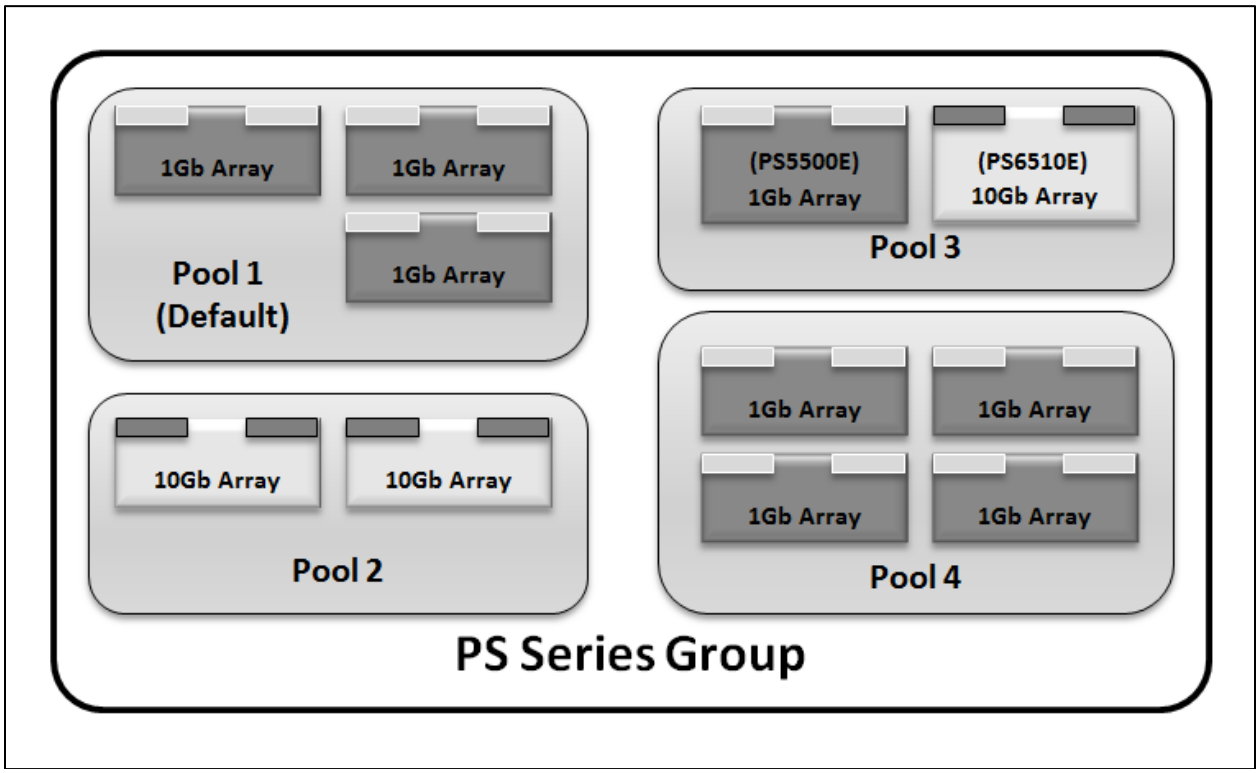


Figure 10 Tiered Pools in a PS Series Group

The following best practices should be considered for storage pools unless the automatic performance load balancer (APLB) is disabled:

- Mixing arrays of different drive speeds and types is fully supported and is the primary reason APLB was created.
- Do not mix arrays with different controller speeds (1GbE, 10GbE) within a single pool unless they are each running unique RAID policies.
- To override the automated performance decisions for a specific volume, indicate a "preferred" RAID type for that volume. If that RAID type exists within the pool, the volume will reside on those arrays that match the preferred RAID policy for the volume.

If the APLB is disabled, the following recommendations apply:

- Do not mix arrays with different drive speeds within a single pool unless they are running a unique RAID policy

- Do not mix arrays with different drive technologies (SATA, SAS, SSD) within a single pool unless they are running a unique RAID policy.
- Do not mix arrays with different controller speeds (1GbE, 10GbE) within a single pool unless they are each running unique RAID policies.

3.3 Volumes

Volumes provide the storage allocation structure within an EqualLogic SAN. Volumes are seen on the network as iSCSI targets by hosts and are presented to the user as disk drives once the iSCSI initiator has been configured and authenticated to the volume. Only computers with an iSCSI initiator and the correct access credentials can access a volume. Disk space for volumes is allocated from the target storage pool's free space.

Volumes are identified within the Group Manager with a unique volume name. The volume name is created by the administrator during volume creation and can be configured with several additional characteristics and capabilities. The following sections provide additional details.

3.3.1 Volume Attributes

Volumes are created using the "create volume" function in Group Manager or through the Command Line Interface (CLI). Each volume must have a unique name that is used by Group Manager to identify and manage the volume. Volume names may be changed via Group Manager at any time. Volume names must meet the following requirements:

- 1 to 64 alpha-numeric characters
- A-Z, a-z, 0-9, ".", "-", ":" are legal characters

Volumes are assigned a unique iSCSI Qualified Name (iqn) that consists of the following parts:

- "**iqn**" followed by a "**.**" (period)
- Year and Month of first full month that the naming authority was registered with standards body. EqualLogic's value is "**2001-5**" followed by a "**.**"
- The storage provider's domain name in reverse order. For example: "**com.equallogic**"
- A colon ("**:**")
- Vendor specified information to identify an iSCSI target or initiator as unique within the vendor domain. For EqualLogic iSCSI targets, this part consists of a unique set of numbers and the user assigned volume name (within Group Manager).

The following is an example of an iSCSI target name for a volume named db3:

```
iqn.2001-05.com.equallogic:7-8b0900-6d0000000-001ebbc5d80sf0k0-db3
```

Volumes are visible to iSCSI initiators through one or more of three mechanisms during volume creation:

- iSCSI initiator IP address
- A host's iSCSI iqn well-known name
- Mutual CHAP authenticated user identity

Note: IQN names are assigned to volumes automatically when they are created. They cannot be changed for the life of the volume. If a volume name is changed, the IQN name associated with the volume will remain unchanged.

3.3.2 Volume Features

Each volume supports a set of features and capabilities that include the following:

- Ability to define a volume as thin-provisioned
- Support for snapshots
- Support for replication
- Support for creating clones
- Support for multiple host shared access

Note: These block level volume features are not applicable to the component volumes created by the NAS Service for the FS Series NAS Reserve.

3.4 Thin Provisioning

Thin provisioned volumes provide administrators with an additional option for managing data growth within the SAN. Thin provisioning allows volumes to present a logical capacity to the host(s) connecting to the volume that is different (larger) than the actual physical storage resource allocation used by the volume at any given time.

Note: Volumes automatically created and managed by the NAS Service inside the NAS Reserve pool used by an FS Series appliance are fully allocated at creation and cannot be thin provisioned. Once those volumes are created (when the NAS Reserve is created they cannot be shrunk.

A volume can be provisioned as a thin provisioned volume at creation or after creation. The following rules apply to thin provisioned volumes:

- A minimum physical allocation of 10% of the logical allocation is required
- If a volume is converted to a thin provisioned volume, physical allocation cannot be less than the amount of physical space already used within the volume
- Any pool free space allocated to a thin provisioned volume is not returned to the free pool if the host's file system usage of that volume is reduced (due to file system defragmentation, data removal, etc.)
- If a thin provisioned volume's allocated space exceeds the "maximum in-use space" setting, the volume will go into an offline state.
- Converting a volume to a thin provisioned volume may not reduce the physical allocation of space the volume is using. Actual physical space recovered will depend on previous writing patterns for the OS file system hosted on the volume.
- All initial settings for "minimum volume reserve", "In-use volume reserve warning level" and "maximum in-use space" are default recommendations and can be changed by the administrator within the constraints defined above.

The following best practices should be considered when using thin provisioned volumes:

- Use Pool Free Space, not Group Free Space when making all determinations of thin provisioned volume physical capacity allocation.
- Create regular volumes before creating thin provisioned volumes. This provides the administrator with a better view of the remaining available free space in the pool.
- Set each thin provisioned volume's "In use Volume Reserve Warning Level" to a level that allows the administrator to ensure that additional physical storage can be deployed before the volume uses the remainder of current pool free space.
- Ensure that the Sum of all "In use Volume Reserve Warning Levels" for all thin provisioned volumes does not exceed the current pool free space capacity minus 10%.
- The storage administrator and application owner should reach an agreement on storage use and procedures in determining the best automatic growth policy, reducing the frequency of monitoring and intervention.
- Be aware of an application's data usage patterns. Some applications perform automated disk optimization that may cause a thin provisioned volume to use more physical storage than would be needed for normal operations. For these types of applications, thin provisioned volumes may not be indicated.
- Use "quick format" options when formatting OS file systems that are hosted by thin provisioned volumes.
- Thin provisioned volumes should not be used for hosting OS boot partitions or OS page file cache.

3.5 Snapshots

Snapshots are point in time copies of volumes. Snapshots have some features and properties similar to volumes as well as some unique capabilities. Like volumes, snapshots can be assigned an iqn and presented as volumes to a host. This allows hosts to mount a snapshot, potentially change the data in the snapshot, or convert it to a real volume (clone) that has dedicated space allocated to it from the free storage pool.

Snapshots require that space be reserved during volume creation (or after volume is created) to store the snapshot(s) created from that volume. All snapshot reserve space for a specific volume always resides in the same storage pool as the volume. By default, Group Manager allocates snapshot reserve space equal to 100% of the host volume's allocated space. This ensures that a 100% data change in the volume can be protected by a single snapshot. This value can be set to a lower value based on the application's data change rate, snapshot/backup plan, or role that the snapshot will be used. The following items should be considered when determine the size of a volume's snapshot reserve:

- Data change rate for the application(s) who is using the volume
- Defragmentation policy of the OS or application using the volume
- The role or purpose for creating the snapshot(s).

Snapshots have the following characteristics:

- They are identified using names that consist of the parent volume name plus a date/time stamp indicating when the snapshot was created as a default name.
- If name of parent volume changes, existing snapshots iqn names do NOT change accordingly, but retain their original iqn name.
- Deleting a snapshot's parent volume deletes all associated snapshots for that volume.

- Snapshots of volumes with a high data change rate will require a larger snapshot reserve space.
- Snapshots have access control lists that are inherited from the parent volume by default.
- Snapshot reserve space for any volume can be decreased at any time. The minimum size allowed will be based on the current space usage consumed by existing snapshots using the snapshot reserve.
- Snapshot reserved space for any volume can be increased at any time assuming there is available free space in the storage pool hosting the volume.

3.6 Clones

Cloning creates a new volume by copying an existing volume. The new volume has the same reported size, contents and thin-provision settings as the original volume. You can clone a regular volume, a specific replica of a volume, or a specific snapshot of a volume.

- Volume clones use non-shared storage (unlike snapshots)
- Volume clones reduce the amount of storage pool free space
- All regular volume operations can be performed on the clone once the cloning operation is complete.
- The cloning operation supports instant accessibility (while the clone operation is occurring)

Table 12 shows clone sources and properties.

Table 12 Cloning Modes and Properties

Clone Source	Common Properties
Clone from Volume	• Creates a new volume • Creates a new volume name • Creates a new iSCSI target IQN • Has same reported size as source volume • Has same contents as source volume • Cloned volume is immediately available • Cloning consumes free pool space equivalent to 100% of the volume reserve setting for the source volume[a]
Clone from Snapshot	
Clone from Inbound Replica[a] (secondary group)	

(a) Replica clones are created in the secondary group pool and are immediately available at the secondary group IP address.

3.7 Template volumes and thin clones

Template volumes and thin clones are new PS Series features introduced with firmware version 5.0. Thin clones are created from template volumes. With a few exceptions, all normal volume operations apply to template volumes and thin clones. For details on using template volumes and thin clones, see the *Advanced volume operations → About template volumes and thin clones* section of the *PS Series Group Manager* documentation.

3.8 Replication

Replication is a powerful feature that can help you manage and implement a disaster recovery strategy for your business applications and data. By replicating business-critical volumes, you ensure that your business operations can resume quickly on a partner group in the event of a disaster on the primary group. You also have the ability to restore the configuration to its original state if the problem on the original group can be corrected. The replication process does not require placing volumes offline. Thus you can take advantage of the benefits of replication without causing interruption in application or data availability.

In EqualLogic PS Series storage, replication is the process of copying volume data from one PS Series group to another PS Series group. Physical distance between the replication groups is not a concern as long as they are connected by a reliable TCP/IP network path.

Table 13 provides definitions of EqualLogic PS Series terminology related to replication processing.

Table 13 Replication Terminology

Term	Description
Replica	A time synchronized copy of an EqualLogic volume stored in a Secondary Group.
Replica Set	A collection of all time synchronized replicas for a specific source volume.
Primary Group	A group containing the source volume(s) to be copied or replicated.
Source Group	Same as "Primary Group".
Secondary Group	A group containing the replica or copy of the source volume(s).
Destination Group	Same as "Secondary Group".
Delegated Space	The amount of space on the secondary group that is "delegated" to a replication partner, to be reserved for retaining replicas.
Replica Reserve	The space allocated from delegated space in the secondary group to store the volume replica set for a specific volume.
Local Reserve	The amount of space reserved on the local or Primary Group for holding temporary snapshots and failback snapshots of the source volume.

3.8.1 Replication Limits

You need to consider the following limits when designing a volume replication topology:

- A primary group can replicate to multiple partner (secondary) replica groups.
- PS series groups can have up to 16 replication partners and can support a maximum of 10,000 total snapshots and replicas from all of its replication partners.
- A group can have volumes replicating with multiple partners, but *an individual volume can have only one replication partner*.
- A maximum of 256 volumes per group can be configured for active replication.
- All volumes that are part of a volume collection can only replicate with a single partner.
- A volume can have up to a maximum of 512 replicas stored on a partner group.
    - For a PS4000 only group, you cannot exceed two replication partners and 32 volumes configured for replication

3.8.2 Replication Paths

The example replication paths shown in Figure 11 are described below.

- **Basic partnership:** one partner group hosts the primary copy of the volume and a second partner group hosts the replica copy of the same volume. We also show the reverse direction of the path for the Fast Failback replication feature, if it is enabled for the replica set.
- **Single to multiple group partnership:** a single group can support replication of multiple primary volumes to different secondary replication partner groups.
- **Reciprocal partnership:** you can create bi-directional replication paths between partner groups
- **Multiple to single group partnership:** a common scenario in which multiple primary groups replicate their volumes to a single secondary group partner in a different location.

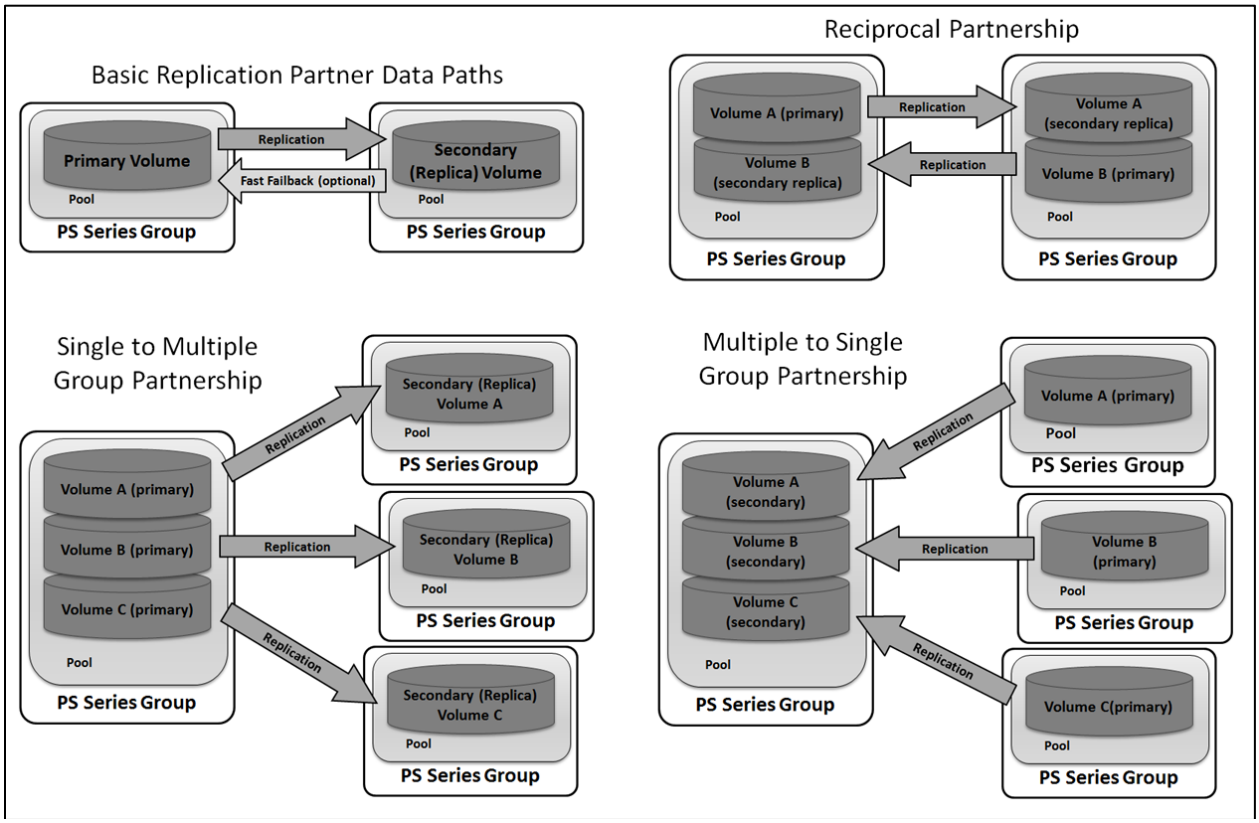
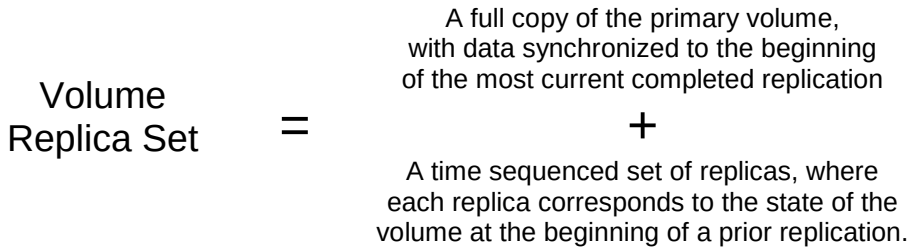


Figure 11 Replication Partnership Paths

3.8.3 Replication Process

When a replica is created the first replication process completes transfer of all volume data. For subsequent replicas, only the data that changed between the start time of the previous replication cycle and the start time of the new replication cycle is transferred to the secondary group. Dedicated volume snapshots are created and deleted in the background as necessary to facilitate the replication process. A volume replication set can be described as a combination of the following:

A volume replica set is defined as follows:


Volume
Replica Set
=
A full copy of the primary volume,
with data synchronized to the beginning
of the most current completed replication
+
A time sequenced set of replicas, where
each replica corresponds to the state of the
volume at the beginning of a prior replication.

The number of prior replicas in the replica set that can stored on the secondary group is limited by the size of the Replica Reserve allocated for that volume and the amount of data that changes.

Replication processing occurs in a series of phases. The flowchart in Figure 12 shows the process phases, focusing on how the process tracks and copies changes that occur between each replica cycle.

Referring to the phases shown in Figure 12:

Replication Setup (one-time):
Configure replication partnership and volume replication settings.

Replication Processing (repeating):
The primary group checks for availability of sufficient delegated and replica reserve space on the secondary group at the beginning of each replication processing phase. If adequate space is not available then the process will pause and generate an event message. Replication will continue once sufficient space is made available. This part of the process is not shown in the chart.

Primary to secondary volume data replication is completed. The process steps vary based on replication status (first or subsequent) and fast failback mode (enabled or disabled). During this process Local Reserve is consumed by a hidden snapshot (and the fast failback snapshot if enabled). Volume data changes that occur during the replication processing phase are stored by the hidden snapshot in Local Reserve. Replica Reserve allocated to the volume within delegated space on the secondary group receives all volume data changes. Replica Reserve is consumed by the most recent complete volume replica plus all prior replicas stored in the replica set.

Between Replication Events (repeating):
Once first replication has occurred the system continues to keep track of volume data changes that occur so that subsequent replication processes can copy those changes to the replica set. This tracking process does not consume additional space.

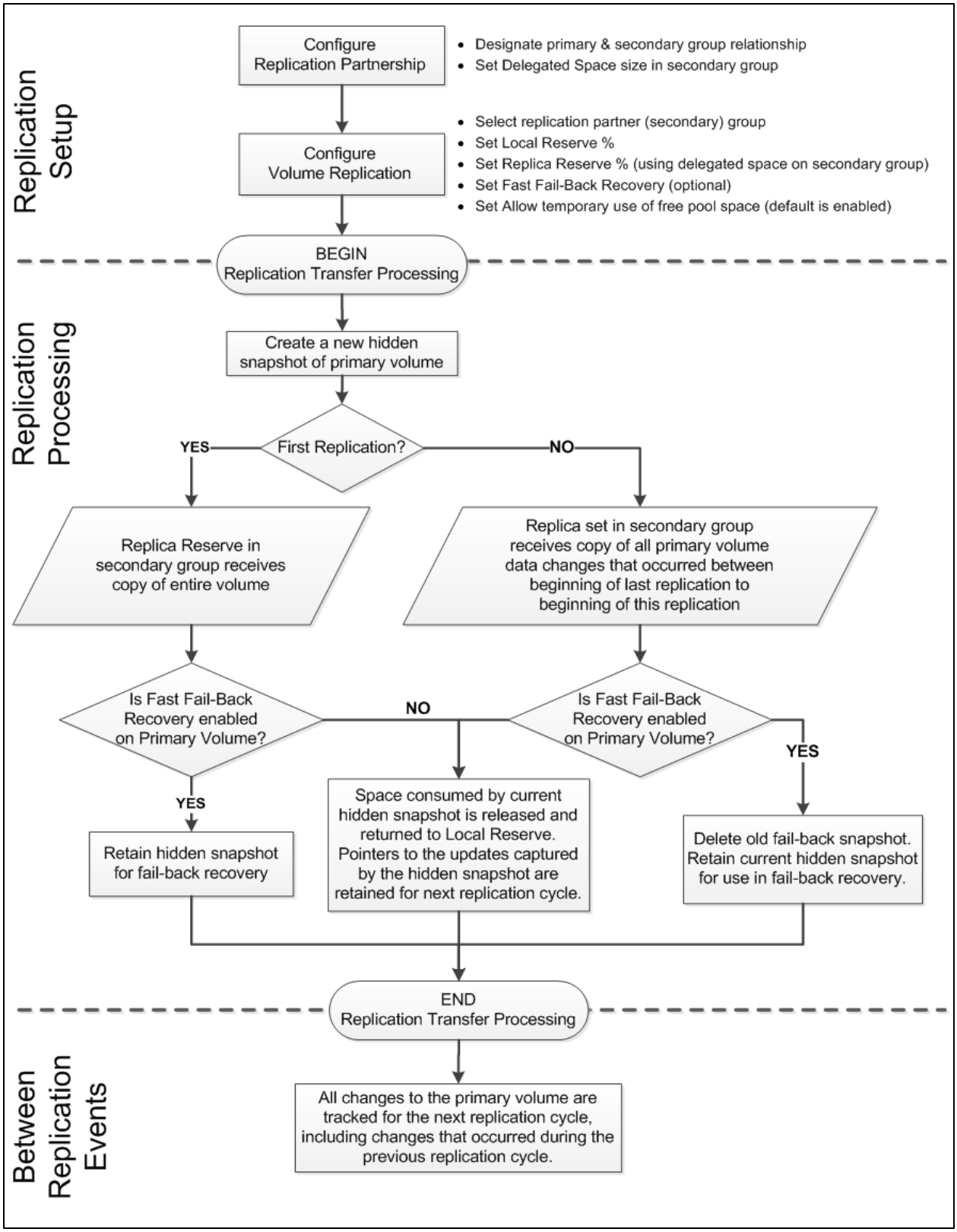

Replication Setup
Configure Replication Partnership
Designate primary & secondary group relationship
Set Delegated Space size in secondary group
Configure Volume Replication
Select replication partner (secondary) group
Set Local Reserve %
Set Replica Reserve % (using delegated space on secondary group)
Set Fast Fail-Back Recovery (optional)
Set Allow temporary use of free pool space (default is enabled)
BEGIN Replication Transfer Processing
Replication Processing
Create a new hidden snapshot of primary volume
First Replication?
YES
NO
Replica Reserve in secondary group receives copy of entire volume
Replica set in secondary group receives copy of all primary volume data changes that occurred between beginning of last replication to beginning of this replication
Is Fast Fail-Back Recovery enabled on Primary Volume?
NO
Is Fast Fail-Back Recovery enabled on Primary Volume?
YES
YES
Retain hidden snapshot for fail-back recovery
Space consumed by current hidden snapshot is released and returned to Local Reserve. Pointers to the updates captured by the hidden snapshot are retained for next replication cycle.
Delete old fail-back snapshot. Retain current hidden snapshot for use in fail-back recovery.
END Replication Transfer Processing
Between Replication Events
All changes to the primary volume are tracked for the next replication cycle, including changes that occurred during the previous replication cycle.

Figure 12 Replication Process

With replication, it does not matter if the volume is thin-provisioned or uses a traditional volume since in either case only the data that has changed will be copied to the replica. On the secondary side, volumes are always thin-provisioned to conserve available capacity used by the Replica Reserve for that volume.

3.8.4 Fast failback

With fast failback enabled you can ensure that the volume data preserved by the failback snapshot on the primary group always matches the volume data in the most recent replica stored on the secondary group. If you have an event that causes failover to the secondary group and the workload subsequently writes changes to the replica volume then failback snapshot supports a quicker fail back to the primary group by replicating only the changes made to the replica volume during the time it was active as a recovery volume on the secondary group. If the failback snapshot is not enabled, you must replicate the entire volume contents back to the primary group to complete the failback operation. Depending on the size of the volume, the failback scenario can take significantly longer to complete if Fast Failback is not enabled.

3.8.5 Sizing Replica Reserve and Delegated Space

The volume replication process in EqualLogic PS Series storage consumes extra storage space on both the primary and secondary group partners. In addition to the space consumed by the volume itself, each volume will require additional space in the primary group for Replication Reserve and Local Reserve, plus delegated space for storing replica sets in the secondary group. This is illustrated in Figure 13 below. A single delegated space on the secondary group must be used for all volumes received from a specific primary group.

Delegated space for a replication partnership must be assigned to a pool. Delegated space for different replication partnerships can be assigned to different pools on the secondary group.

Note: Any type of application or operating system level process that causes block level data changes will affect replication data size and time to complete. For example, file system defragmentation of a large partition will significantly increase the amount of space consumed in replica reserve, delegated space on the secondary group and time to complete the next scheduled replication.

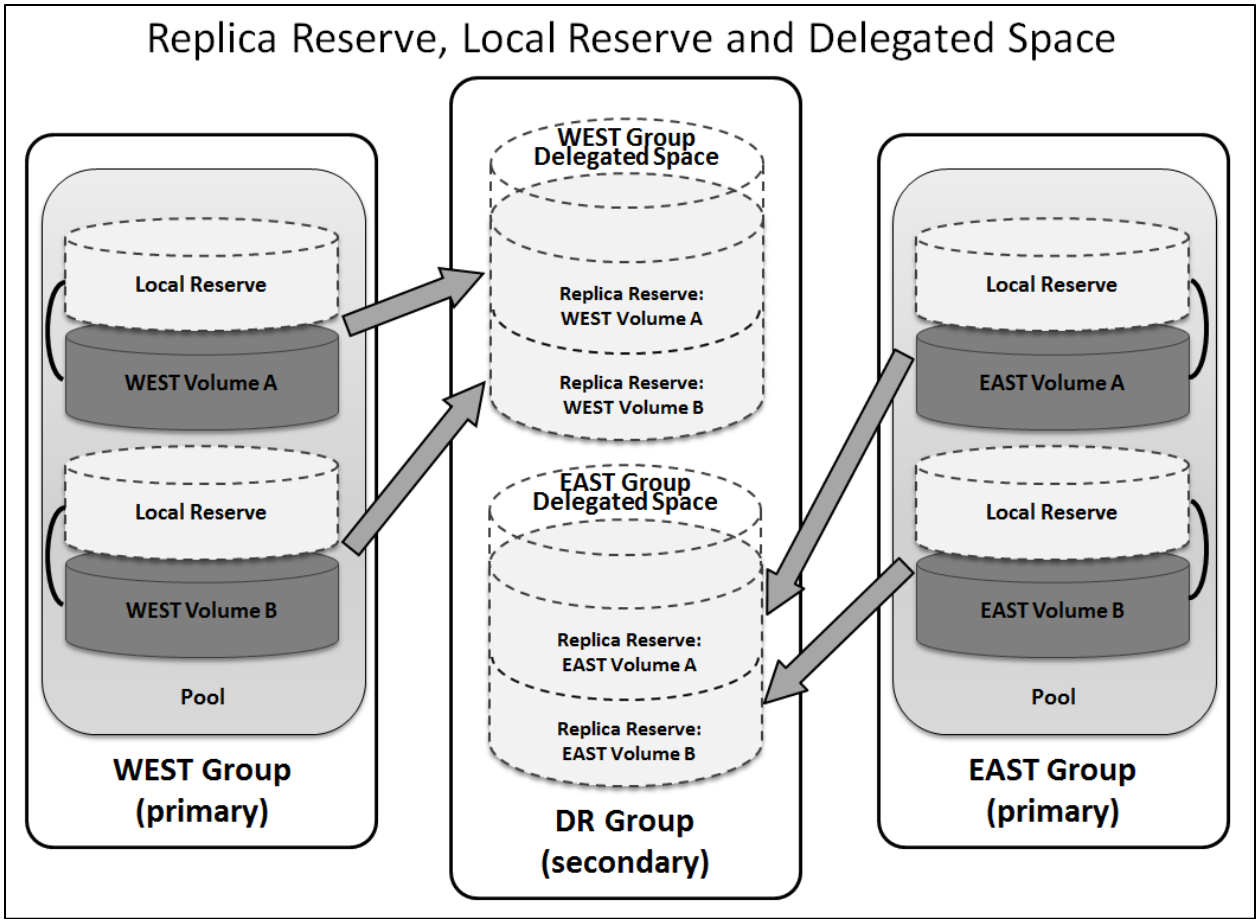


Figure 13 Replica Reserve, Local Reserve, and Delegated Space

Auto-replication requires reserved disk space on both the primary and secondary groups. The amount of space required depends on several factors:

- Volume size.
- The amount of data that changes (on the source volume) between each replication period.
- The number of replicas that need to be retained on the secondary site.
- If a failback snapshot is retained on the primary group.

The default values that appear in Group Manager are sufficient to ensure that enough space is reserved for at least one successful replication, even if the entire contents of a volume are altered between replicas. Initial replication of a volume will cause the complete contents of the volume to be copied to the secondary group. Subsequent replication events will copy only the changed data. The recommended and space efficient guidelines for sizing replication reserves and delegated space are presented in Table 14.

Table 14 Replication Space Sizing Guidelines

Replication Space	Recommended Value		Space Efficient Value
Local Reserve (Located on Primary Group.)	No Failback Snapshot:	**100%**	**5% + CR[(a)]**
	Keep Failback Snapshot:	**200%**	**10% + CR[(a)]**
Replica Reserve (Located on Secondary Group.)	**200%** Ensures there is adequate space for the last replica and any replica in progress.		**105%[(b)]** **+ CR x (# of Replicas – 1)**
Delegated Space (Located on Secondary Group. The replica reserve space for all replica sets coming from a single group)	Must be large enough to hold the sum of all replica reserve sizes for all volumes replicating to that group.		Monitor change rate, adjust to lower than recommended value and continue monitoring.

(a) CR: "Change Rate". For details on how volume changes affect replication space, see the Centralized Replication section under Replication Configuration Options in the Group Manager Help documentation.
(b) Start with 105%, then add to that the maximum number of replicas expected to be stored in the replica set minus 1, multiplied by the expected Change Rate.

3.8.6 Effect of TCP Window Size on Latency across WAN links

In certain cases, increasing the TCP Window Size setting on PS Series controllers can improve replication throughput across WAN links. The ability to change this setting is currently not supported using the EqualLogic Group Manager. Customers interested in changing TCP Window size settings should contact Dell EqualLogic Technical Support for assistance. For more details on how this setting can affect replication performance across WAN links, see the following publication:

- *Dell EqualLogic Auto-Replication: Best Practices and Sizing Guide*
  http://www.delltechcenter.com/page/Dell+EqualLogic+Auto-Replication+Best+Practices+and+Sizing+Guide

3.8.7 Replication Partner Compatibility

Table 14 provides details on replication partner interoperability based on array firmware version. For the latest information, consult the release notes for your array firmware version.

Table 15 Firmware Replication Compatibility

Firmware of "Local" Group	Firmware on Replication Partner
V5.1.x, V5.2.x	V5.0.x, V5.1.x, V5.2.x
V5.0.x	V4.2.x, V4.3.x, V5.0.x, V5.1.x, V5.2.x
V4.2.x, V4.3.x	V4.1.x, V4.2.x, V4.3.x, V5.0.x
V4.1.x	V4.0.x, V4.1.x, V4.2.x, 4.3.x
V4.0.x	V3.2.x, V3.3.x, V4.0.x, V4.1.x, V4.2.x
V3.2.x, V3.3.x	V3.0.x, V3.1.x, V3.2.x, V3.3.x, V4.0.x
V3.0.x, V3.1.x	V3.0.x, V3.1.x, V3.2.x, V3.3.x

3.9 Clustering

To support a shared storage environment, EqualLogic allows concurrent access to any volume. Concurrent volume access is enabled on a per volume basis within the Group Manager or via CLI. The EqualLogic array will not manage concurrent access to these volumes. Control of access to a shared volume must be provided by means of access control software on each host that has access to the volume.

Warning: Failure to provide access control to a shared volume can result in data loss.

By default, EqualLogic PS Series groups disable multi-host (shared) access to target.

- If needed, you can enable multi-host access for a volume.
- If you enable multi-host access to a volume, then proper I/O operation concurrency must be maintained.

4 EqualLogic SAN Design

An EqualLogic iSCSI SAN can be operated in any network that supports the industry standards and IP subnet design guidelines described in this section. Because of this flexibility, there are many network design and configuration choices that can affect SAN performance. The following sections provide details related to network design and configuration to support the use of an EqualLogic SAN.

Note: With the addition of FS Series NAS functionality, the EqualLogic product family now provides an iSCSI based **unified file and block storage platform**. An EqualLogic SAN can now be used to provide **block level access** (direct iSCSI access to PS Series arrays) or **file system level access** (via the FS Series appliance) using NFS or CIFS protocols and the Dell FluidFS scale-out file system. Unless otherwise stated, recommendations in this document are applicable to both file and block environments.

4.1 Implementation of Standards

EqualLogic SANs are based on industry standards. The following standards are required to support all host to target communications and member to member communications in an EqualLogic SAN:

- IETF Standards
  - IETF RFC1122 "Requirements for Internet Hosts – Communications Layers"
  - IETF RFC1123 "Requirements for Internet Hosts – Application and Support"
  - IETF RFC3270 "Internet Small Computer Systems Interface (iSCSI)"
- IEEE Standards
  - 802.1
  - 802.3

4.1.1.1 iSNS Support

An Internet Storage Name Service[1] (iSNS) server can support discovery, management and configuration of group members by dynamically updating information about the iSCSI target names for group volumes. Once the IP address of an iSNS server is entered in an iSCSI initiator's configuration utility, the setting is persistent across initiator sessions. A PS Series group can be configured to register with up to three iSNS servers.

Note: Starting with Firmware V4.1.4, volume and snapshot identifiers are no longer automatically published to iSNS servers. This applies to new volumes and snapshots as well as volumes and snapshots that existed before the group was upgraded to V4.1.4.

4.2 General Requirements and Recommendations

For EqualLogic PS Arrays, the following general SAN design ***requirements*** apply:

- For all members (arrays) in a given SAN Group all ports should be connected to the same subnet. This allows the arrays to communicate with each other as a group of peer members. The arrays must be in the same subnet as the group's "well known" IP address.

[1] The Internet Storage Name Service (iSNS) specification: http://tools.ietf.org/html/rfc4171

Note: Hosts can be in a different subnet as long as those hosts have layer 3 routing available to the subnet containing the arrays and the group's well known IP address.

- Rapid Spanning Tree Protocol must be enabled if the SAN infrastructure has more than two switches in a non-stacked configuration, and portfast must be enabled on all edge device ports (hosts, FS Series appliances and arrays).
- Port density requirements to support fully redundant configurations and maximum SAN throughput are as follows:
    - **PS4x00 family**: 2x 1GbE ports per controller = 4x 1GbE ports total
    - **PS4110 family**: 1x 10GbE port per controller = 2x 10GbE ports total
    - **PS5x00 family**: 3x 1GbE ports per controller = 6x 1GbE ports total
    - **PS6x00 family**: 4x 1GbE ports per controller = 8x 1GbE ports total
    - **PS6510 family**: 2x 10GbE ports per controller = 4x 10GbE ports total
    - **PS6110 family**: 1x 10GbE port per controller = 2x 10GbE ports total
    - **FS7500 NAS**: 12x1GbE ports per controller node (four client LAN and eight iSCSI SAN) + 1 100/1000Mb port per controller node for IPMI interconnection
- At least two iSCSI SAN ports per host (block level iSCSI access) are required for fully redundant SAN connectivity. Host ports can be 1GbE or 10GbE.
- Quality of Service (QoS) based on what is traditionally designated as IEEE 802.1p is not currently supported for use with EqualLogic SANs. QoS and Class of Service designations must be disabled.
- All switches within the SAN must be interconnected such that there is always a path from any Ethernet port on one array to all other Ethernet ports on all other arrays in the group.
- All switches and host network controllers within the infrastructure must have flow control enabled for optimal performance.
- Any EqualLogic SAN group that is required to send or receive replication traffic to/from another SAN group must have an uninterrupted communications path (ie. "visibility") between each group.
- To prevent a switch failure from also disabling all paths between a host and its connected volumes, all ports from each controller need to be connected to at least two different switches. For PS4100/PS6100 family arrays, split the vertical port pair connections between two switches to ensure 100% bandwidth capability is maintained in the event of a vertical port failover event. See the configuration diagrams in Section 4.5.1.
- The above guideline regarding path redundancy across multiple switches also applies in FS Series NAS appliance configurations. (Refer to the connection diagrams in Section 7.1)
- **For SANs connecting to an FS Series appliance, all switches in the SAN must have jumbo frames enabled.**

For EqualLogic PS Series Arrays, the following general SAN design *recommendations* apply:

- Take advantage of your switch's VLAN capabilities. You should create a VLAN dedicated to iSCSI traffic (even on dedicated switches). If necessary, create a second VLAN for management traffic. The actual VLAN configuration of your iSCSI SAN will be dictated by your SAN network design requirements and the features of the iSCSI SAN switches being used.
- Jumbo frames should be enabled. If you choose to use jumbo frames then all nodes in the SAN fabric must have jumbo frames enabled.

- For best performance and reliability, we recommend that all interconnection paths between non-stacking switches (LAGs) use a dynamic link aggregation protocol such as LACP

4.2.1 Quality of Service (QoS)

Quality of service is described as either of the following:

- The ability to provide different priority levels to different applications, users, or data flows, or to guarantee a certain level of performance to a data flow.
- A network function implemented in some routers and switches that provides high priority for certain types of latency-sensitive traffic (for example, VoIP) and lower priority for other types of traffic (for example, web or http).

PS Series arrays are designed to provide I/O as fast as your network can support it. Therefore, using QoS with iSCSI traffic does not produce expected or desirable results on the SAN. Also, QoS rules can affect how well–or even whether–replication succeeds between PS Series groups. If you plan to use QoS, Dell recommends that you use it only on VLANs that do not carry iSCSI traffic, or on WANs, where bandwidth is shared with other applications and the PS Series array uses it for time-insensitive replication traffic.

Dell recommends against using QoS on the SAN.

4.3 Ethernet Switches and Infrastructure

Any switch used in an EqualLogic SAN should meet the requirements listed in this section.

Note: A detailed and frequently updated list of recommended switches is maintained in a separate document:
Validated Components List for EqualLogic PS Series SANs
http://www.delltechcenter.com/page/EqualLogic+Validated+Components

Note: The FS7500 NAS appliance requires the use of 1Gb switches that meet the requirements in this section, *and provide ports that can support dual 1000/100Mbs link speeds*.

An EqualLogic SAN consists of one or more hosts connected to one or more PS Series arrays through a switched Ethernet network. To support a high performance Dell EqualLogic SAN, switches must meet the following general requirements:

- **Low latency:** Switches with relatively high latency may cause SAN throughput performance to degrade, and under high load conditions they could increase the risk of dropped connections.
- **Non-blocking backplane design:** SAN Switches should be able to provide the same amount of backplane bandwidth to support full duplex communication on ALL ports simultaneously.
- **Adequate buffer space per switch port:** In addition to supporting data transfers between the hosts and the SAN, Dell EqualLogic arrays also use the SAN to support inter-array communication and data load balancing. For this reason, the more buffer space per port that a switch can provide the better. Due to the multitude of buffer implementations used by switch vendors, Dell cannot provide definitive guidelines as to how much is enough. Port buffers should be designed such that data is not lost when traffic reaches extreme levels. Due to the

clustered storage traffic patterns used by the EqualLogic SAN architecture, switches that support "cut-through" mode are not suitable for use in an EqualLogic SAN and may actually result in lower overall SAN performance.

- **Support for IEEE 802.3x flow control (passive and/or active) on ALL ports:** Switches and network interface controllers used in an EqualLogic SAN must be able to passively respond to any "pause" frames received. If possible, you should use switches that have the ability to transmit "pause" frames to external devices in the event that the device cannot adequately forward traffic in a timely fashion.
- **Support for Jumbo Frames:** This is not a requirement. But, the use of jumbo frames may yield desirable results. Most iSCSI SAN implementations should benefit from using jumbo frames. The actual impact on SAN throughput when using jumbo frames will depend on your workload's I/O characteristics.
- **Support for Rapid Spanning Tree protocol (IEEE 802.1w), or Cisco "portfast" functionality if the SAN infrastructure will consist of more than two switches:** For SAN infrastructures consisting of more than 2 non-stacking switches, R-STP must be enabled on all ports used for inter-switch trunks. All non-inter-switch trunk ports should be marked as "edge" ports or set to "portfast".
- **Support for unicast storm control:** iSCSI in general, and Dell EqualLogic SANs in particular can send packets in a very "bursty" profile that many switches could misdiagnose as a virally induced packet storm. Since the SAN should be isolated from general Ethernet traffic, the possibility of actual viral packet storms occurring is non-existent. In an EqualLogic SAN, the switches must always pass Ethernet packets regardless of traffic patterns.
- **Support for Inter-Switch Trunking (IST) or Stacking:** IST support is required to link all switches in SAN infrastructure together. For stacking capable switches, the use of stacking ports for IST is assumed. A good rule of thumb for stacking link bandwidth would be a minimum 20 Gbps full-duplex.
- **Support for vLAN functionality if SAN traffic is to share the same physical switch resources with other (non-iSCSI SAN) network traffic.**

Note: We recommend this strategy for small SAN environments only, where networking device resources are limited. For larger SANs and datacenter environments we recommend using vLAN functions only if Data Center Bridging (DCB) is implemented.

- **Support for creating Link Aggregation Groups (LAG):** For non-stacking switches, the ability to bind multiple physical ports into a single logical link for use as an inter-switch trunk (IST) is required. The switch should support designating one or more ports for IST (via Link Aggregation Groups). The switch should support creation of LAGs consisting of at least eight 1Gbps ports or at least two 10Gbps ports.

Note: For 1GbE SANs, using non-stacking switches to connect three or more EqualLogic arrays into a single group may negatively impact SAN I/O throughput performance.

4.3.1 Connecting SAN Switches in a Layer 2 Network

When more than one SAN switch is required, each switch connected to the array group members will be in the same subnet. These switches must be interconnected to provide a single switched Ethernet

fabric. Figure 14 shows the two common methods for interconnecting switches, using either stacking switches or non-stacking switches.

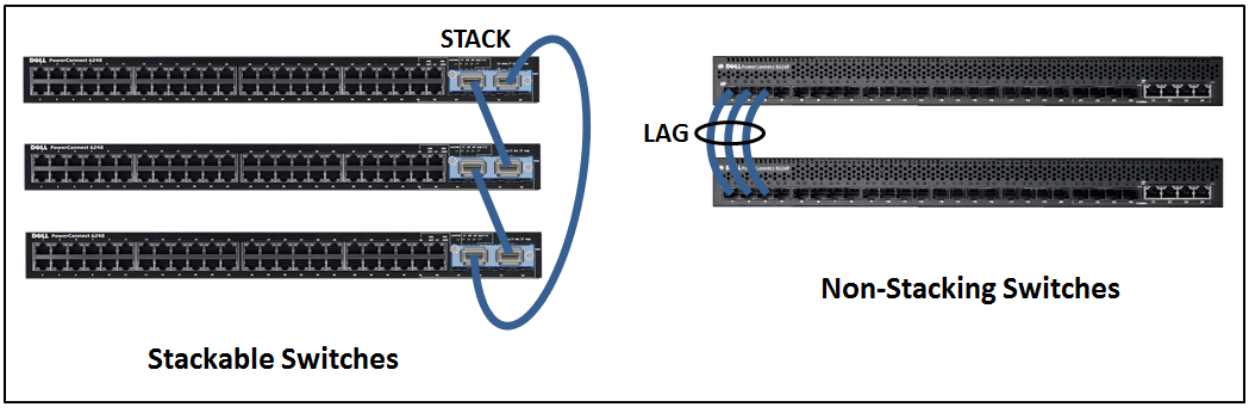


Figure 14 Switch Interconnects

4.3.1.1 Stacking Switches

Stacking switches provide the preferred method for creating an inter-switch connection within a Layer 2 network infrastructure. Stacking is typically accomplished using a vendor proprietary, high-bandwidth, low-latency interconnect that allows two or more switches to be connected in such a way that each switch becomes part of a larger, virtual switch. A stackable switch will provide a set of dedicated stacking ports. Installation of an optional stacking module may be required. Considerations for stacking link bandwidth:

1Gb Switches	The stacking link bandwidth should be at least 10Gb/s in each direction on each wire (full-duplex) to provide adequate throughput to support an EqualLogic SAN consisting of 1Gb arrays.
10Gb Switches	The stacking link bandwidth should be at least 40Gb/s in each direction on each wire (full-duplex) to provide adequate throughput to support an EqualLogic SAN consisting of 10Gb arrays or a mix of 1Gb and 10Gb arrays.

4.3.1.2 Non-Stacking Switches

Non-stacking switches do not have a dedicated switch interconnect. In this case to create an interconnect between the switches you must utilize one or more ports on each switch to create a Link Aggregation Group (LAG). This type of Inter-switch connection should utilize link aggregation functions (if provided by the switch) to aggregate multiple Ethernet ports into a single, logical high bandwidth interconnect. There are several options depending on the vendor of the switch. Table 16 describes the most common options.

Note: For best performance and reliability, we recommend that all interconnection paths between non-stacking switches (LAGs) use a dynamic link aggregation protocol such as LACP.

Table 16 Link Aggregation Types

Link Aggregation Type	Notes
Static	Static link aggregation defines a set of links that provide a point to point connection between two switches. These links may or may not provide failover redundancy or traffic load management.
LACP	Link Aggregation Control Protocol is based on IEEE 802.3ad or IEEE 802.1AX. LACP is a dynamic LAG technology that automatically adjusts to the appearance or disappearance of links within the defined LACP group.
PAgP	Port Aggregation Protocol (PAgP) is a Cisco Systems® proprietary networking protocol, which is used for the automated, logical aggregation of Ethernet switch ports, known as an etherchannel. This means it can only be used between Cisco switches and/or switches from licensed vendors.
Vendor Proprietary	Several switch vendors may provide additional link aggregation options that are completely proprietary or may be extensions to one of the two previously mentioned solutions. In most cases, this type of link aggregation solution is designed to reduce or eliminate the need – and the overhead – of the Spanning Tree Protocol that must be used in the two previous options. If available, these proprietary options should be considered. They may be very useful in allowing the network administrator to create a more efficient multi-switch layer 2 network infrastructure for a SAN. Be aware that these proprietary solutions must support the transmission of IEEE 802.1x flow control and jumbo frames (if used) to properly support an EqualLogic SAN.

4.3.1.3 Using a LAG to Connect Stacked Switches

In some situations it may become necessary to expand the EqualLogic SAN by using more than one single switch stack. For example, you can link multiple stacks by creating a multi-link LACP based LAG between the switch stacks. A simplified stack plus LAG switch configuration is illustrated in Figure 15. You should consider the following recommendations when designing this type of SAN:

- If possible, use 10Gb connections for all links between each stack.
- Distribute the links across multiple switches within each stack (this is known as a "cross-stack" link aggregation group).
- Use LACP or another type of dynamic link aggregation protocol.
- Perform tests to determine the best hashing algorithm to use within the link aggregation group (port channel).

Note: A multi-stack SAN infrastructure as described in this section may provide added reliability to the SAN environment. But, it may also introduce additional latency and the potential for lower throughput. The SAN designer will have to carefully consider the performance and reliability implications.

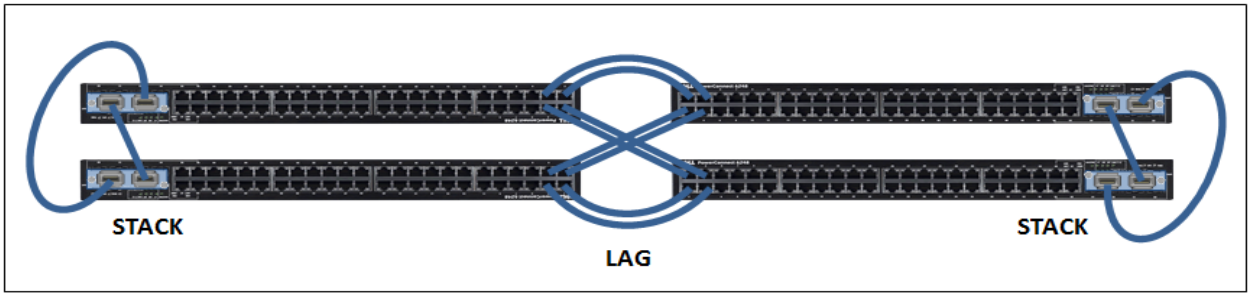


Figure 15 Using a LAG to Interconnect Switch Stacks

4.3.2 Sizing Inter-Switch Connections

Use the guidelines in Table 17 as a starting point for estimating Inter-switch connection sizes.

Table 17 Switch Interconnect Design Guidelines

Connection Speeds	Interconnection Guidelines
1GbE switches attached to 1GbE array controllers	**1-5 arrays:** 1Gb of IST bandwidth per active array controller port (up to the aggregated maximum bandwidth of the IST. **6+ arrays:** Use 1-5 array rule, then add 1Gb of additional bandwidth for each array added
10GbE switches attached to 10GbE array controllers	**PS6010/PS6510 or PS4110/6110 (Random Small Block Workloads):** **1-5 arrays**: 20 – 30Gb of IST bandwidth between each switch **6+ arrays:** At least 6Gb of IST bandwidth per array between each switch **PS4110/PS6110 (Large Block Sequential Workloads):** **1-5 arrays:** 10Gb of bandwidth per active array controller port (up to the aggregated maximum bandwidth of the IST). **6+ arrays:** Use 1-5 array rule, then add 10Gb of additional bandwidth for each array added
1GbE switches connecting to 10GbE switches in a mixed speed SAN	Actual requirements will vary. For more details see Section 4.6, *Integrating 1GbE and 10GbE SANs*

4.3.3 Comparing Inter-Switch Connection Types

Table 18 provides details and recommendations for selecting interconnect options.

Table 18 Stacking vs. Inter-Switch Trunking

Interconnect Type	Primary Purpose	Analysis
Stacking	Create a larger, logical switch within an isolated physical location.	Advantages: • Easier to manage multiple switches as single switch • Higher bandwidth than using link aggregation and Ethernet • Not limited by Ethernet standards Concerns: • Proprietary, cannot be used to interconnect switches from different vendors • Increases cost of switch • Stack bandwidth overload risk Recommendation: • Best way to scale a storage network in a single location on a single subnet • Provides lower latency and higher bandwidth than inter-switch trunking • Understand the stacking technology and limit stack size in accordance with total throughput requirements and connection counts.
Link Aggregation Groups (LAG)	Create a data path between switches in one location or subnet with those in another location or subnet	Advantages: • Leverages Ethernet standard extensions • Can be used to interconnect switches from different vendors • Can use Link Aggregation Protocols (LACP/EtherChannel) to pool multiple 1GbE or 10GbE links into a single logical link providing bandwidth and redundancy Concerns: • Most solutions limited to 8 port link aggregation group • Spanning Tree Protocol must be used if more than two switches are used causing some links to be "blocked" reducing bandwidth availability • LAG bandwidth overload risk Recommendation: • Use when stacking is not available • Use when connecting to aggregation/core switching infrastructure • Use when switches are from different vendors

4.4 Block I/O Host Configuration Guidelines

Designing a redundant SAN requires the availability of redundant NICs or HBAs on each server. A redundant NIC configuration on the server requires at least two NICs installed into separate PCI slots in the server. Table 19 below shows how to achieve redundant server NIC connection configurations for a server with three installed NICs.

Table 19 Redundant Server NIC Configurations

	LOM NIC	Installed NIC 1	Installed NIC 2
NIC Connections to SAN	X	X	-
	-	X	X
	X	-	X

4.4.1 Design guidelines for host connectivity in a redundant SAN

Using the Dell PowerEdge R610 server as an example, you configure redundant connection paths to the SAN switches as shown in Figure 16 below. The R610 server shown in Figure 16 has one additional dual-port PCI-E NIC installed. This configuration leaves two unused ports on the LOM controller for connecting to the server LAN.

Note: As a best practice, we recommend using the same NIC devices in your server for all connection paths to the iSCSI SAN. This will minimize the complexity of server configuration management.
For maximum performance, ensure that the PCI express slot hosting the network controller has the same specifications as the network controller. For example, if the network controller has a PCIe x8 interface then ensure that it is installed into a PCIe slot that can support 8 or more PCIe lanes.

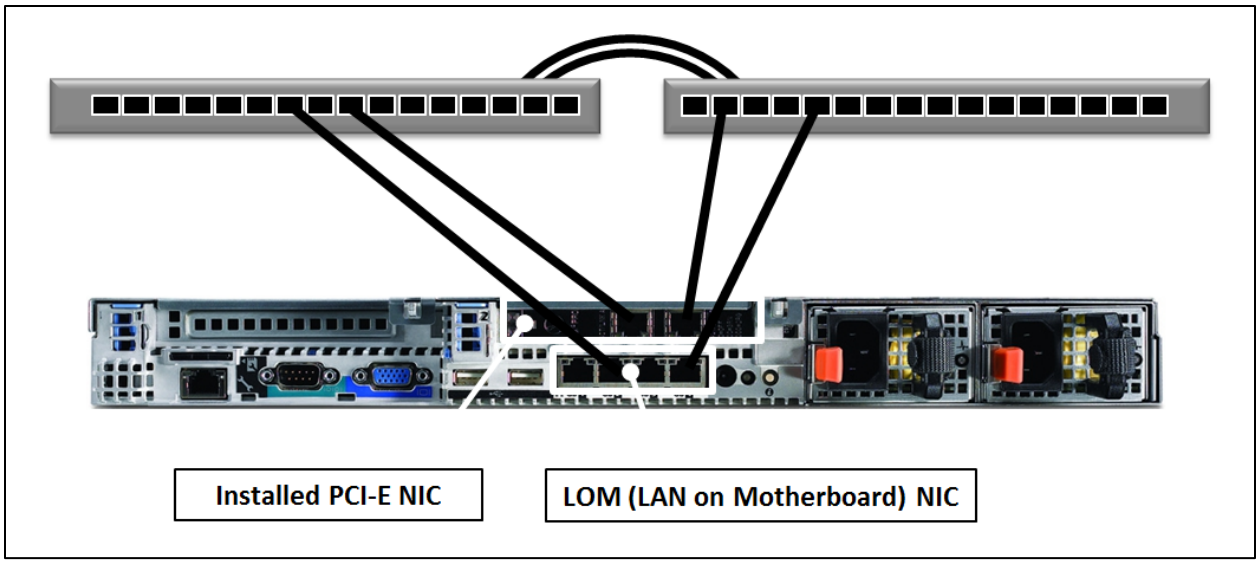


Figure 16 Redundant NIC Connections from Server to SAN using one installed PCI-E NIC and LOM

An R610 server with two additional dual-port PCI-E NICs installed is shown in Figure 17 below. This configuration leaves all four ports on the LOM controller available for other connections.

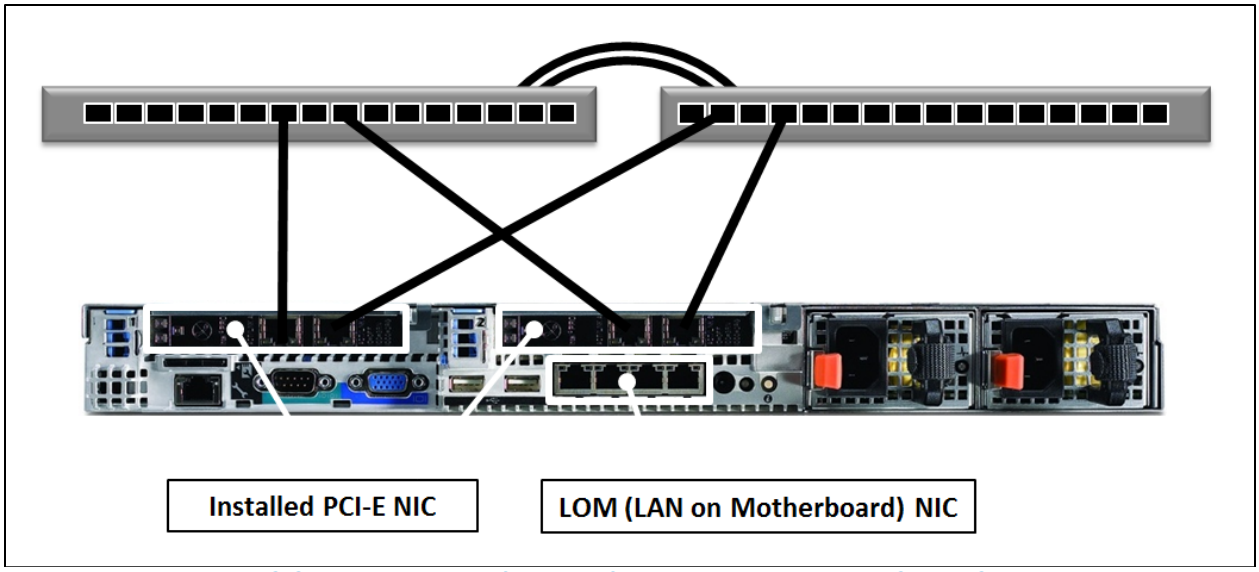


Figure 17 Redundant NIC Connections from Server to SAN using two installed PCI-E NICs

4.4.2 Multi-Path I/O

There are generally two types of multi-path access methods for communicating from a host to an external device. For general networking communications, the preferred method of redundant connections is the teaming of multiple NICs into a single, virtual network connection entity. For storage, the preferred method of redundant connection is the use of Multi-Path IO (MPIO). Though some storage solution can and do support either method for iSCSI connectivity, EqualLogic requires the use of MPIO to enable multiple NIC/HBA connections to be utilized for access to an EqualLogic SAN.

4.4.2.1 EqualLogic MPIO Requirements

The following host port requirements must be met to use MPIO with EqualLogic SANs:

- At least two (2) Ethernet ports are required on each host.
- The host operating system must have a supported MPIO driver or service available.
- The ports used for MPIO cannot be "teamed" to other ports.
- The ports must be the same speed
- The ports must be assigned IP addresses on the same subnet

4.4.2.2 EqualLogic MPIO General Recommendations

Follow this general set of guidelines for configuring MPIO on a host:

- Configure volume access controls to use standard iSCSI IQN names (See Section 3.3.1 for details). For a more secure configuration you can use the IQN name plus the CHAP security ID.
- On each array enable at least two(2) ports for host connectivity.
- Install the Dell provided MPIO extension features if available for the host operating system.
- For Microsoft Windows, install the Device Specific Module (DSM) found in the Host Integration Toolkit for Windows.
- For VMware vSphere 4.1, install the EqualLogic Multipathing Extension Module.
- For other operating systems, use the native MPIO functionality.

4.4.2.3 Configuring Microsoft Windows MPIO

Configure Microsoft Windows MPIO with the following initial configuration settings. Customized settings may be required depending on the supported application(s).

- Change the "Subnets included" field to include ONLY the subnet(s) dedicated to the SAN network infrastructure.
- Change the "Subnets excluded" field to include all other subnets.
- The "Load balancing policy" should remain set to the default value of "Least queue depth".
- "Max Sessions per volume slice" should be set to the number of network ports dedicated to SAN subnet (maximum of 4).
- "Max sessions per entire volume" should be set to three (3) times the value of "Max sessions per volume slice" (maximum of 12).
- "Use MPIO for snapshots" should remain at the default setting.
- "Use IPv6 or IPv4" should be set to IPv4 unless your network is configured to use IPv6 as the default communications protocol.

4.4.2.4 Configuring VMware vSphere 4.1 Multipathing Extension Module (MEM)

Configure the vSphere MEM with the following initial configuration settings. Customized settings may be required depending on the supported application(s).

- Change the "Subnets included" field to include ONLY the subnet(s) dedicated to the SAN network infrastructure.
- "membersessions" should be set to the number of network ports dedicated to SAN subnet (maximum of 4).
- "volumesessions" should be left at the default value of 6.
- "totalsessions" should be left at the default value of 512.

4.5 EqualLogic iSCSI SAN Design

This section will combine all of the SAN components and information provided so far to present a redundant EqualLogic PS Series SAN design. We also include a series of examples illustrating partially redundant and non-redundant SAN configurations.

Note: For FS Series NAS design, the only SAN design patterns presented in this section that we recommend in support of the NAS Service are the fully redundant and fully cabled examples shown in Section 4.5.1. Fully cabled and redundant connection diagrams for the FS7500 are shown in shown in Section 7.1.

The information provided here does not address all of the possible variations in a customer network environment. All information is presented using a set of basic reference designs that make the following assumptions:

- The SAN network is physically isolated from all other network traffic
- The examples are based on best practice design principles.
- Unless otherwise stated, all reference designs will provide end-to-end host to volume redundant paths

- A minimal number of switches will be illustrated to allow for the design concept to be understood. Actual implementations will vary depending on your network infrastructure requirements.
- If sharing physical switches with other non-SAN traffic, we assume all switches are VLAN capable.

4.5.1 Redundant SAN Configuration

In a redundant iSCSI SAN, each component of the SAN infrastructure has a redundant connection or path. The following figures show example connection paths necessary to create a Redundant SAN.

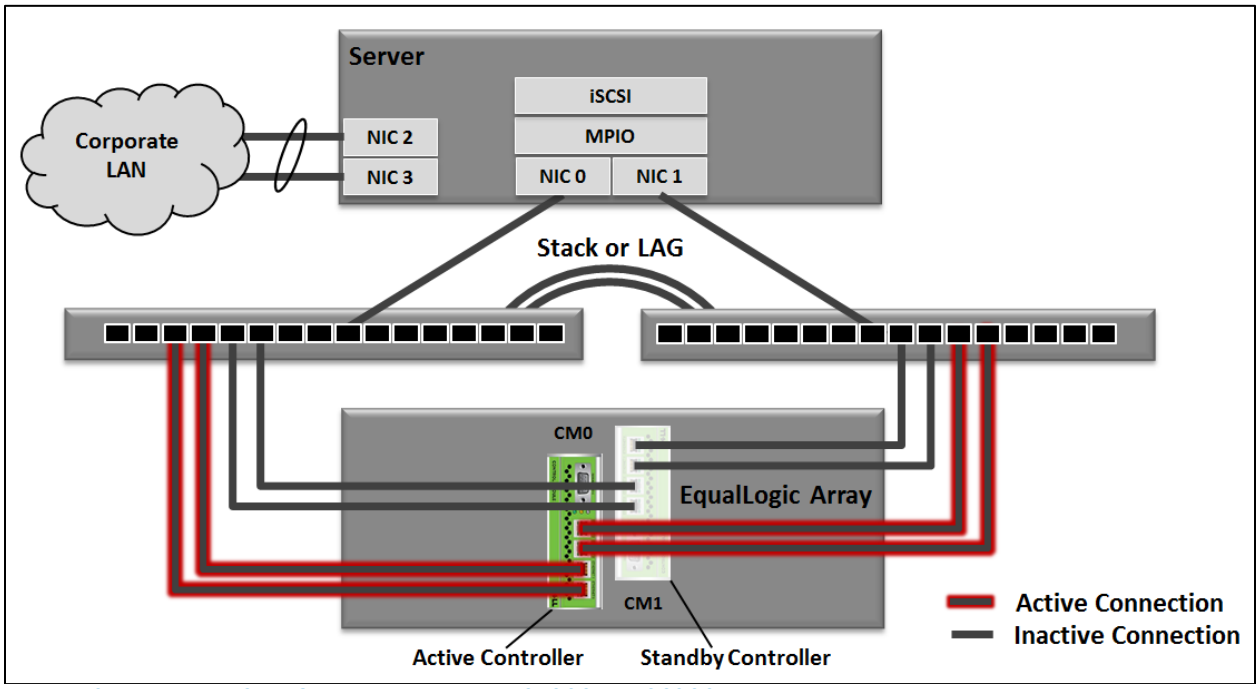


Figure 18 Redundant SAN Connection Paths: PS3000 to PS6000 Family Arrays

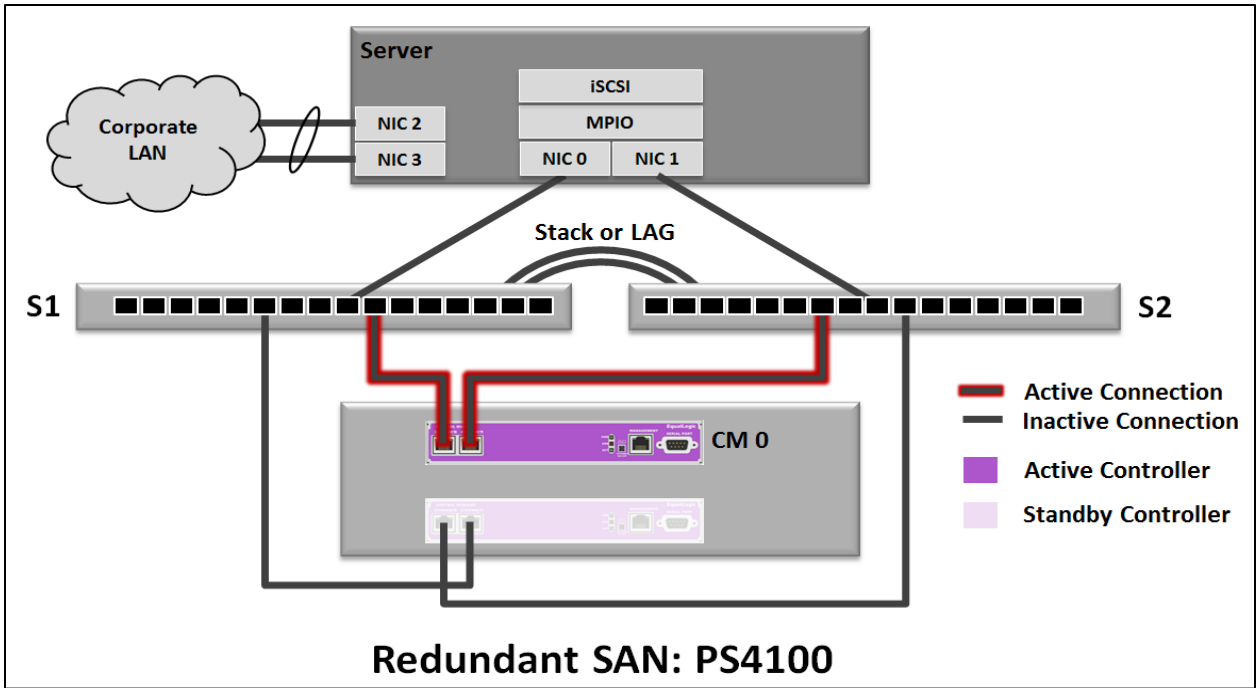


Figure 19 Redundant SAN Connection Paths: PS4100

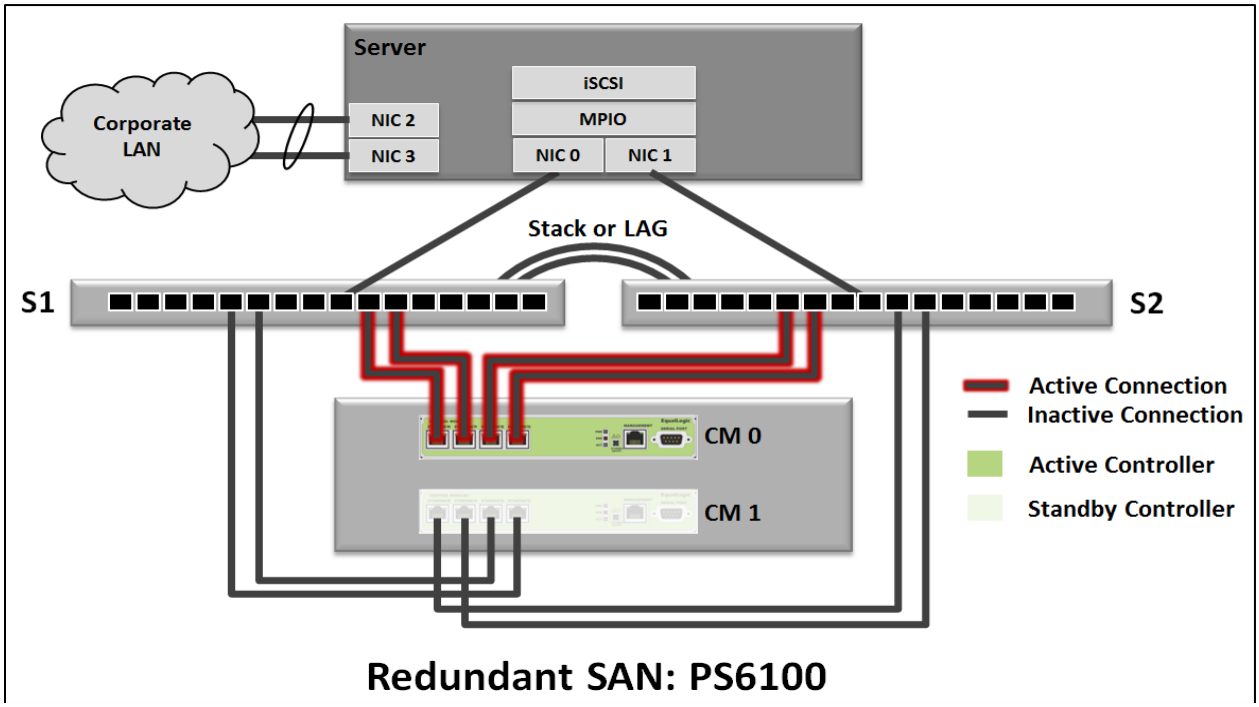


Figure 20 Redundant SAN Connection Paths: PS6100

Note: For a production environment, the configuration examples shown above will protect your access to data. These are the ONLY SAN configurations recommended by Dell.

4.5.2 Partially Redundant SAN Configurations

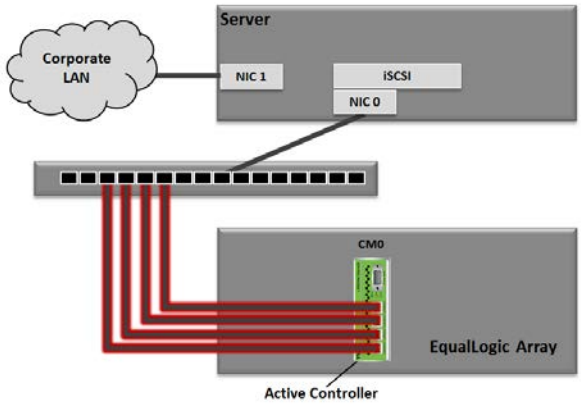
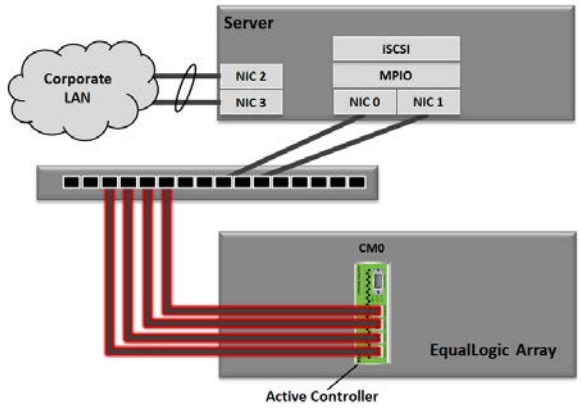
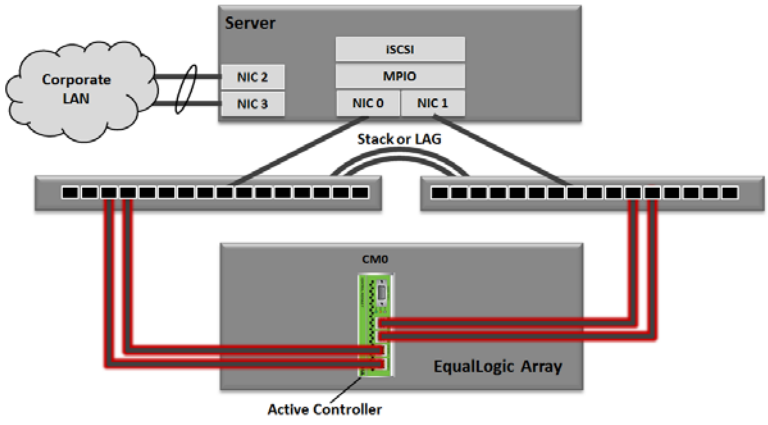
Each of the SAN configurations shown in this section will allow host connectivity to data stored in the SAN. These configurations are for reference only, and the methods shown apply to both PS3000-

PS6000 family controllers and PS4100/PS610 family controllers. **They are not recommended for production deployment since they do not provide end-to-end redundant connection paths.**

4.5.2.1 Single Array Controller Configurations

Table 20 below shows configurations using a single array controller.

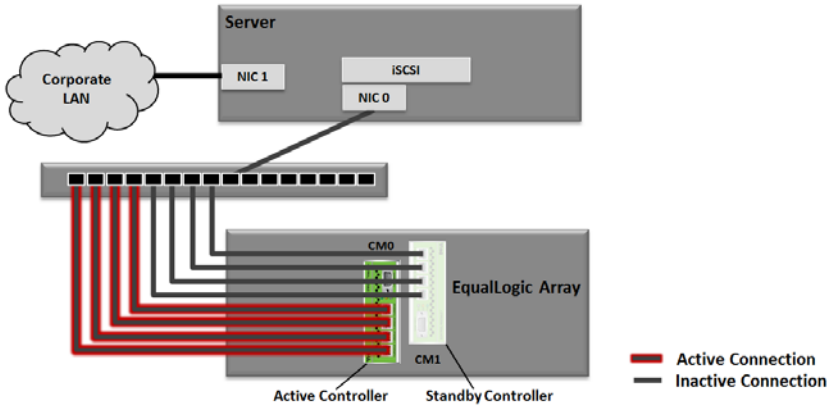
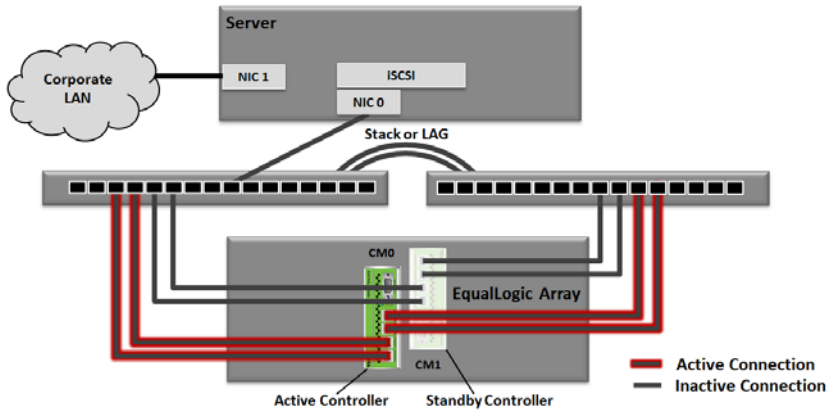
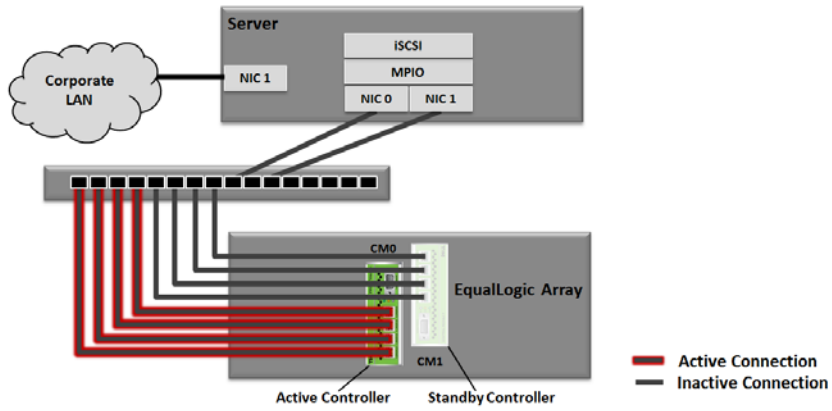
Table 20 Single Controller Array Configurations

Single NIC **Single Switch** **Single Controller**	
Dual NIC **Single Switch** **Single Controller**	
Dual NIC **Dual Switch** **Single Controller**	

4.5.2.2 Dual Array Controller Configurations

You can configure a Dell EqualLogic array to run using dual controllers. Table 21 below shows configurations using a single array controller.

Table 21 Dual Controller Array Configurations

Single NIC **Single Switch** **Dual Controller**	
Single NIC **Dual Switch** **Dual Controller**	
Dual NIC **Single Switch** **Dual Controller**	

4.5.3 Minimum Cabling Scenarios: PS4100 and PS6100

The vertical port failover feature (described in Section 2.3.7) allows you to cable dual controller PS4100 and PS6100 family arrays for maximum I/O bandwidth and controller redundancy while using only one half of the available controller ports. The diagrams in Figure 21 and Figure 22 show how to connect these arrays to accomplish this.

Note: The example configurations shown in this section only apply to the PS4100 and PS6100 family arrays and are recommended only when you do not have available SAN switch ports necessary to support fully cabled configurations.

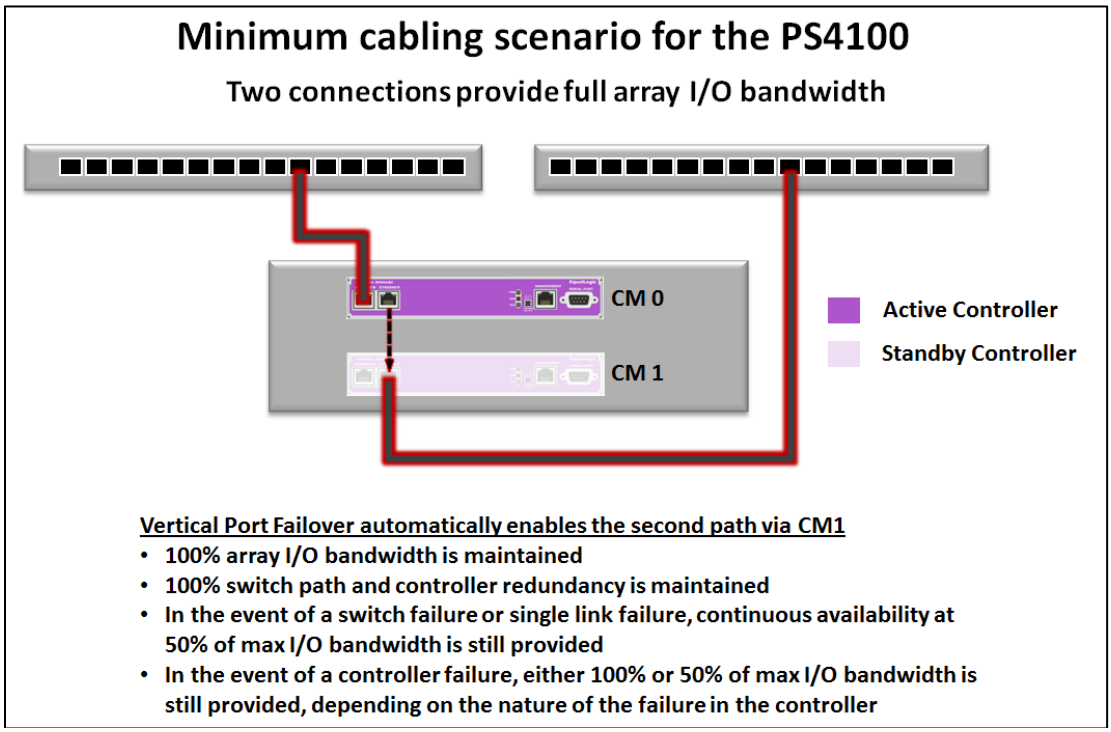

Minimum cabling scenario for the PS4100
Two connections provide full array I/O bandwidth
CM 0
CM 1
Active Controller
Standby Controller
Vertical Port Failover automatically enables the second path via CM1
• 100% array I/O bandwidth is maintained
• 100% switch path and controller redundancy is maintained
• In the event of a switch failure or single link failure, continuous availability at 50% of max I/O bandwidth is still provided
• In the event of a controller failure, either 100% or 50% of max I/O bandwidth is still provided, depending on the nature of the failure in the controller

Figure 21 Minimum Cabling Scenario: PS4100

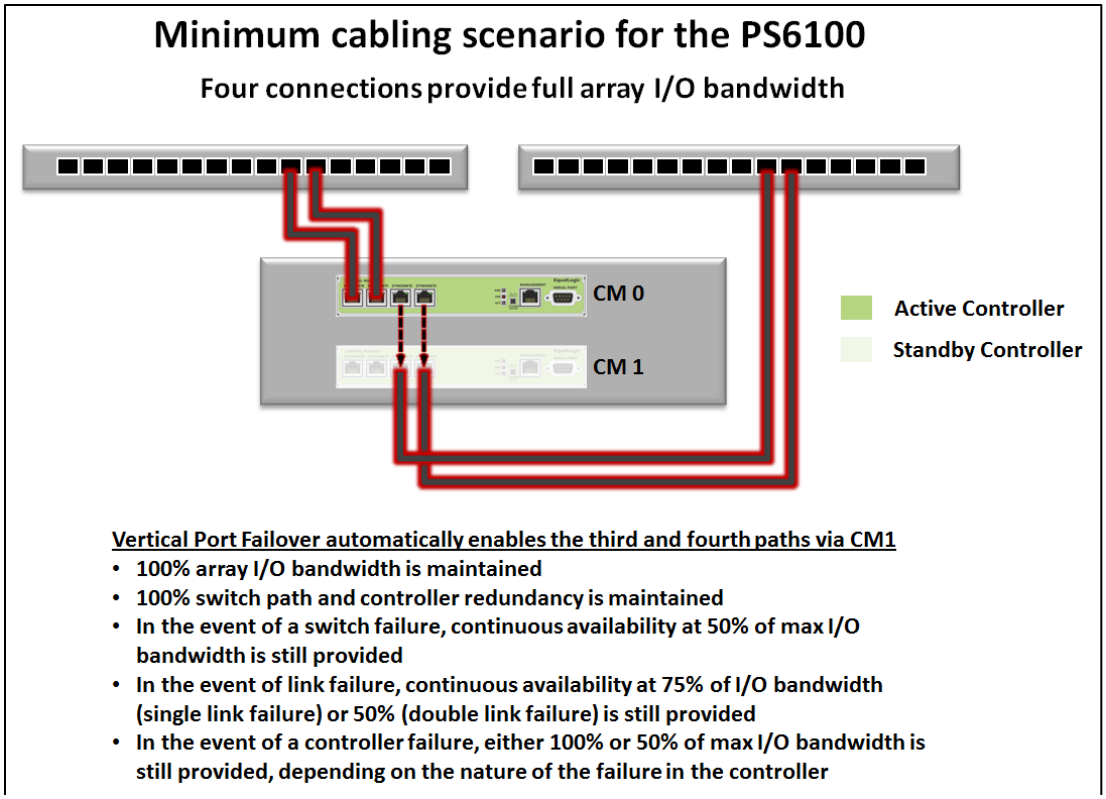

Minimum cabling scenario for the PS6100
Four connections provide full array I/O bandwidth
CM 0
CM 1
Active Controller
Standby Controller
Vertical Port Failover automatically enables the third and fourth paths via CM1
• 100% array I/O bandwidth is maintained
• 100% switch path and controller redundancy is maintained
• In the event of a switch failure, continuous availability at 50% of max I/O bandwidth is still provided
• In the event of link failure, continuous availability at 75% of I/O bandwidth (single link failure) or 50% (double link failure) is still provided
• In the event of a controller failure, either 100% or 50% of max I/O bandwidth is still provided, depending on the nature of the failure in the controller

Figure 22 Minimum Cabling Scenario: PS6100

4.6 Integrating 1GbE and 10GbE SANs

With the introduction of 10GbE, there will be situations that require 1Gb arrays and 10Gb arrays coexisting in the same SAN infrastructure. EqualLogic PS Series arrays support operation of 1Gb and 10Gb arrays within the same group. This section summarizes mixed speed SAN design guidelines that are presented in much more detail in the following publications:

Best Practices for Deploying a Mixed 1 Gb/10 Gb Ethernet SAN using Dell EqualLogic Storage Arrays:
Deploying+Mixed+1+Gb/10+Gb+Ethernet+SANs+using+Dell+EqualLogic+Storage+Arrays

Integrating EqualLogic PS6x10 Arrays with Existing SANs:
http://www.equallogic.com/resourcecenter/assetview.aspx?id=9447

The potential advantages in running a mixed speed (1GbE <-> 10GbE) EqualLogic SAN include:

- Not all of the application workloads on a SAN will require storage I/O performance that the 10Gb arrays provide. Thus, SAN administrators will have additional storage tiering flexibility based on array I/O performance.
- The PS Series Group Manager will allow the SAN administrator to still manage both types of arrays within the same SAN group.
- The ability to mix 1Gb and 10Gb arrays supports seamless operational coexistence during during migration to a 10Gb SAN.

4.6.1 Design considerations

To properly implement a mixed speed SAN, you must pay close attention to the following design and implementation considerations:

- Ethernet switch feature set, port density and required configuration settings
- Optimization of Rapid Spanning Tree Protocol behavior
- Optimal switch interconnect pattern
- Awareness of I/O workload patterns coming from 1Gb and 10Gb initiators vs. target volume locations in the SAN
- I/O performance *implications when using mixed speed vs. segregated speed pools*

To create a redundant, mixed speed iSCSI SAN, at a minimum we recommend that you start with dual 1GbE and dual 10GbE switches. Figure 23 shows an example SAN design, where two switches of each type are used.

Referring to Figure 23:

- The design is based on using features provided by the Dell PowerConnect 6248 1Gb Ethernet switch and the Dell PowerConnect 8024 10Gb Ethernet switch.
- The Dell PowerConnect 8024 is not a stackable switch, so a link aggregation group (LAG) is used to create the inter-switch trunk paths.
- Each of the 1Gb switches is configured with one dual-port 10GbE uplink module and one stacking module. The 10GbE uplink modules are used for creating 20Gb LAG uplinks to the 10Gb switches.
- **Split Interconnect** –The 20Gb LAG uplinks between the 1Gb and 10Gb switches are cross-connected so that each 10Gb switch physically connects to both switches in the 1Gb stack.
- **FS7500 Connection Path** – The initial FS Series NAS product (FS7500) is a 1Gb only solution. The NAS Reserve pool for the FS7500 could also be stored in a mixed 1Gb/10Gb PS Series SAN. In this case the FS Series appliance must connect via the 1Gb switch path as shown in Figure 23 and Figure 24.

The 10Gb switches are connected together using multiple 10Gb links. This provides us a path for devices connected directly to the 10Gb switches to communicate without having to go through the slower 1Gb switches.

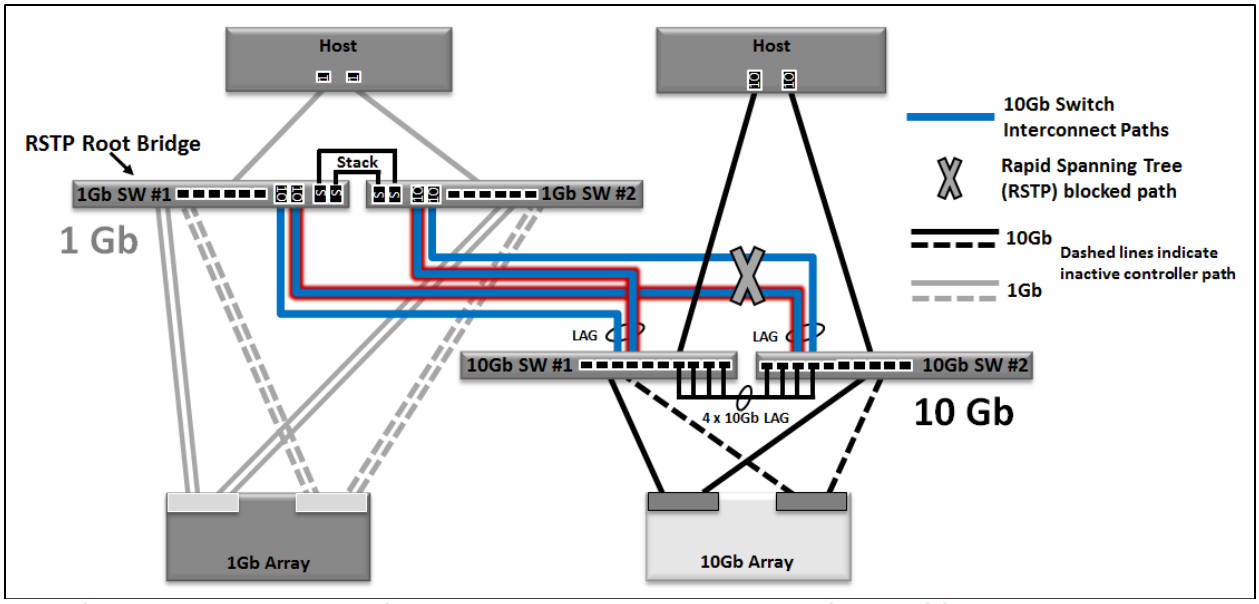


Figure 23 Mixed speed redundant SAN using split interconnect between 1Gb and 10Gb switches

4.6.1.1 Optimizing Rapid Spanning Tree Protocol Behavior

The LAG between the 10Gb switches in Figure 23 creates a loop in the network. Rapid Spanning Tree Protocol (RSTP) will compensate for this by blocking paths as necessary. The optimal spanning tree strategy for this network is to prevent RSTP from blocking the inter-switch trunk between the 10Gb switches, thus causing some 10Gb traffic to traverse the slower 1Gb switches. To accomplish this you need to be aware of which switch is acting as the root bridge in the spanning tree. For the mixed speed SAN design shown in Figure 23 the root bridge is **1Gb SW#1**. Based on this information you can assign a link cost to ensure that the desired link configuration is achieved. Note the location of the RSTP blocked path in Figure 23. We manually assigned a high link cost to one of the 20Gb uplink LAGs so that it became the RSTP blocked path.

4.6.1.2 You must use the straight uplink pattern when using non-stackable switches

If you are using switches that do not support a stacking mode then you must use the straight interconnect uplink pattern shown in Figure 24. Note the following design differences between the split uplink pattern in Figure 23 and the straight uplink pattern in Figure 24:

- A LAG is used to create the connection between **1Gb SW#1** and **1Gb SW#2**.
- A high rapid spanning tree link cost is assigned to the 1Gb switch LAG (note the location of the RSTP blocked path in Figure 24). Doing this prevents 10Gb inter-switch traffic from having to pass through the 1Gb switch LAG.

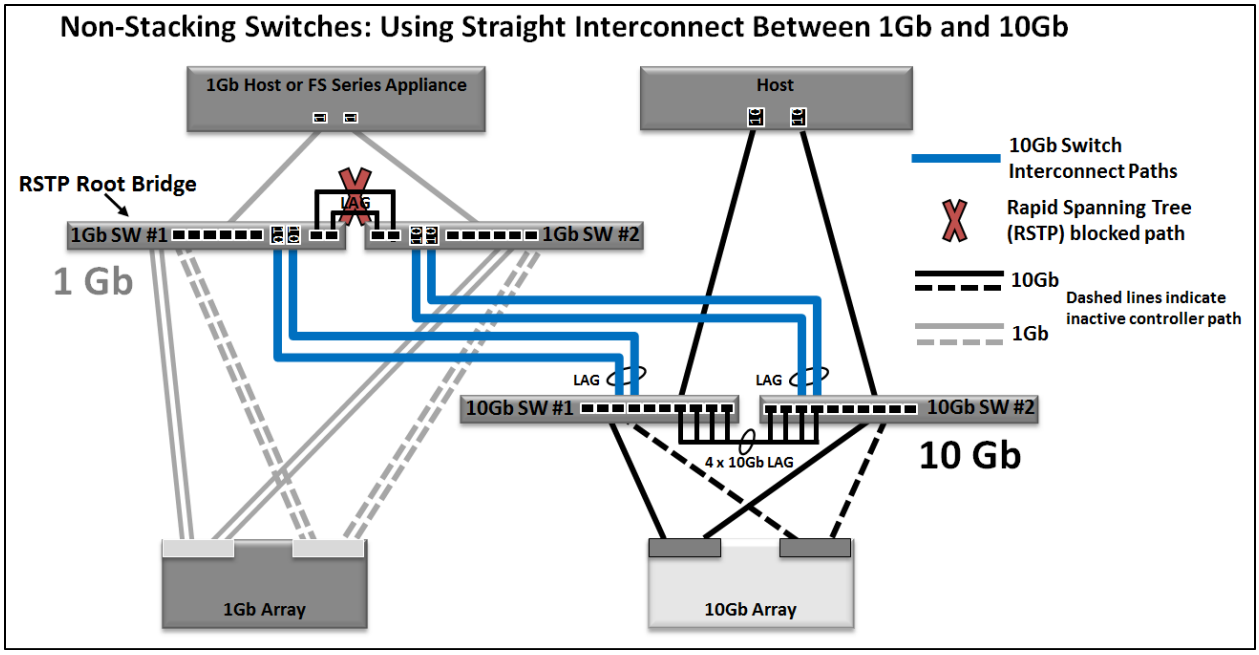


Figure 24 Mixed speed redundant SAN using straight interconnect between 1Gb and 10Gb switches

4.6.2 Mixed SAN best practices

The following list summarizes the important SAN design considerations for integrating 10Gb EqualLogic arrays into existing 1Gb EqualLogic SANs.

- When integrating 10Gb switches into your existing 1Gb environment, how you interconnect the mixed speed switches (split vs. straight uplink) does not have a significant impact on performance as long as the uplinks are sized appropriately to your workloads.
    - If your 1Gb switches are configured as a stack then you should use the split interconnect pattern shown in Figure 23 by default.
    - If your 1Gb switches are not stacked, then you must use the straight interconnect pattern shown in Figure 24.
- When connecting 1Gb switches and 10Gb switches together you must always be aware of where Rapid Spanning Tree is going to block links to make sure that 10Gb traffic (i.e. EqualLogic inter-array data flow) never crosses the 1Gb switch.
- You must configure pools and volumes in a way that minimizes impact to IO performance.
    - If you have predominately 1Gb initiators, start upgrading your arrays to 10Gb for comparable or better performance across almost all situations.
    - If you have predominately 10Gb initiators, you should only access data and volumes residing on 10Gb arrays (from those initiators). You may see high latency and retransmit rates when 10Gb initiators connect to 1Gb targets.
    - When adding 10Gb arrays, place them in separate pools from your 1Gb arrays.

4.7 Blade Server Chassis Integration

Integrating the PowerEdge M1000e Blade Server Solution (or any third party blade chassis implementation) requires additional SAN design considerations. Each M1000e can support up to three separate networking "fabrics" that interconnect ports on each blade server to a pair of blade IO modules within each chassis fabric through an intervening chassis midplane interface. Each fabric is

associated with different interfaces on a given blade server as described in Table 22. Each blade server has a "LAN on Motherboard" capability that is mapped to the IO modules located in the Fabric A IO modules slots on the M1000e chassis and only supports 1Gb or 10Gb Ethernet networking depending on the blade server model. In addition, each blade server has two "mezzanine" sockets for adding additional networking options such as 1Gb or 10Gb Ethernet, Infiniband, or Fibre Channel cards. These mezzanine cards are mapped to either the Fabric B or the Fabric C IO modules.

Figure 25 illustrates the layout of the three fabric IO modules located on the back of the M1000e chassis.

Table 22 M1000e Fabric Mapping

	LOM/MLOM	Mezzanine B	Mezzanine C
Fabric	A	B	C

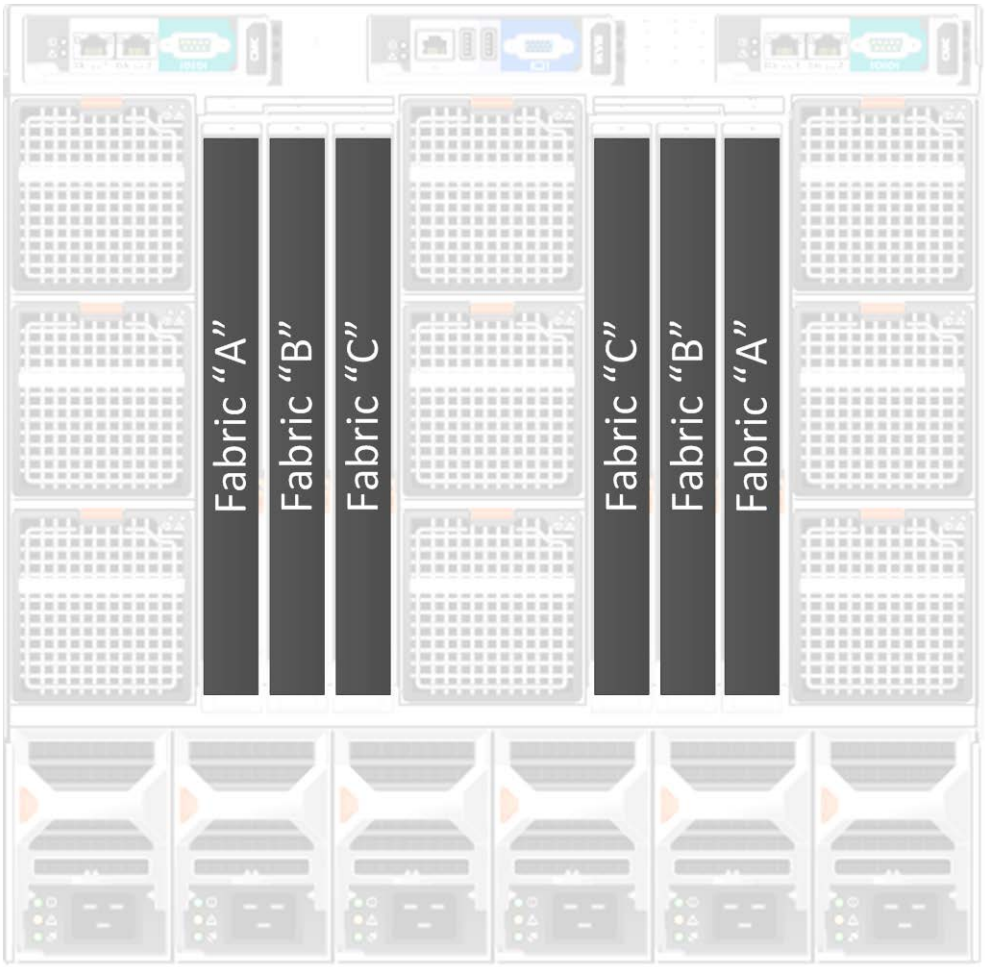


Figure 25 Blade IO Modules and M1000e Chassis

There are two primary methods of integrating blade chassis with EqualLogic PS Series arrays.

- Directly attaching EqualLogic arrays to the blade IO modules on each chassis.

- Creating an external SAN switch infrastructure to host all array connections and using the blade IO modules as "host access" switches.

For each of these two general SAN design strategies, the user must make design decisions with respect to the type of interconnects to use between the blade IO modules and/or the external stand-alone switches. Depending on the design recommendation, stacking, link aggregation, or a combination of both types of interconnect may be used. In Section 4.7.1 we will discuss strategies for directly attaching arrays to blade IO modules within one or more M1000e chassis. In Section 4.7.2 we will discuss strategies for connecting M1000e blade chassis to an external SAN infrastructure.

4.7.1 Designing a SAN using Blade Chassis IO Modules with Directly-Attached Arrays

The following points should be taken into consideration when planning a SAN solution that requires directly attaching EqualLogic arrays to PowerConnect M-Series blade IO modules:

- Limited number of externally accessible ports typically available on blade IO modules

  Current PowerConnect M-Series blade IO modules have limited numbers of externally accessible Ethernet ports. For a 1GbE SAN solutions each array will require up to 4 ports per IO module and for a 10GbE SAN solution each array will require up to 2 ports per IO module. Based on these maximum requirements, the table below provides a breakdown on the number of externally accessible ports on each model of PowerConnect M-Series blade IO modules and the maximum number of arrays that could be directly attached to a single M1000e blade chassis.

Table 23 Blade IO Module Options for EqualLogic

	Maximum Available External Ports	Ports Recommended for Interconnect	Arrays Supportable per M1000e
PowerConnect M6220	8x 1GbE	0(Stackable)	2
PowerConnect M6348	16x 1GbE	0(Stackable)	4
PowerConnect M8428-k	8x 10GbE	2x10GbE(PS6010/6510) 3x10GbE(PS4110/6110)	3(PS6010/6510) 5(PS4110/6110)
PowerConnect M8024-k	8x 10GbE	2x10GbE(PS6010/6510) 3x10GbE(PS4110/6110)	3(PS6010/6510) 5(PS4110/6110)
Cisco Catalyst Blade Switch 3032	8x 1GbE	4x1GbE	1
Cisco Catalyst Blade Switch 3130G	4x 1GbE	0(Stackable)	1
Cisco Catalyst Blade Switch 3130X	4x 1GbE	0(Stackable)	1

- All switches within the SAN infrastructure, including blade IO modules, must provide a path from all ports hosting array connections to all other ports within the SAN.
- An important networking requirement for EqualLogic SAN infrastructure design is that all switches must be part of the same Layer 2 Ethernet fabric if they will be hosting connections from EqualLogic arrays. For this reason, if arrays are going to be directly attached to one or more blade chassis, then the IO modules within the Fabric that will be hosting the arrays must

be interconnected. Since the IO modules are not interconnected via the chassis midplane, the only alternative is to use external ports (stacking or Ethernet) to make these inter-switch connections.

- Maximum stack size versus practical stack size
- Depending on the IO module model, the maximum number of switches allowed in a single stack may be different. In addition, the number of switches supported by the switch may not be optimal for a SAN using EqualLogic arrays. SAN solutions tend to be sensitive to latency and adding multiple hops between a host and target may create situations where latency becomes unacceptable for your application. Testing is advised.
- Each M1000e enclosure will contribute two switches to the stack placing an upward limit on the number of enclosures that can be interconnected into a single SAN network. Regardless of the maximum number of IO modules supported in a single stack, Dell recommends not going above six switches in the stack due to possible hop related latency.
- M-Series IO Module Stacking Compatibility
- Not all M-Series IO modules can be stacked together or with external switches. Typically, each M-Series IO modules model can only stack with modules of the same exact model. It may also be possible to stack M-Series IO modules with some "stack compatible" stand-alone PowerConnect switch models. The table below provides stack compatibility information.

Table 24 Stack Compatibility for M-Series IO Modules

	Stack Compatible Switches	Maximum Stack Size	Recommended Stack Size
PowerConnect M6220	PowerConnect M6220 PowerConnect 6224 PowerConnect 6248	10	4
PowerConnect M6348	PowerConnect M6348 PowerConnect 7024 PowerConnect 7048	12	6
PowerConnect M8428-k	N/A	N/A	N/A
PowerConnect M8024-k	M8024-k PowerConnect 8024 PowerConnect 8024F	6	6
Cisco Catalyst Blade Switch 3032	N/A	N/A	N/A
Cisco Catalyst Blade Switch 3130G	Catalyst BS3130G Catalyst BS3130X	8	4
Cisco Catalyst Blade Switch 3130X	Catalyst BS3130G Catalyst BS3130X	8	4

4.7.1.1 SAN Design for a Single M1000e Enclosure

For a single M1000e enclosure, there is a requirement to use two blade IO modules for redundancy. Depending on the blade IO module model, you may use either stacking or link aggregation as described earlier in Table 23. Where stacking is supported, this will most likely be the primary method of interconnecting the blade IO modules to create a single switch fabric. SAN switches can be located in any of the three enclosure fabrics and it is not required that the blade IO modules be in the same enclosure fabric, though it is typical that the two IO modules would be in the same enclosure fabric.

Placing the blade IO modules in the same fabric does remove one aspect of high availability in that each blade server will have both of the SAN ports located on the same fabric mezzanine card. This creates a potential single point of failure if the mezzanine card as a whole fails.

One alternative configuration would be to place the two blade IO modules into enclosure slots associated with two different enclosure fabrics (B1 and C1 for example). This has the advantage that each blade server will have its Ethernet ports connected to these blade IO modules on two different mezzanine cards. This ensures that even if a single mezzanine card fails, there is still an active port on the SAN network.

Table 25 and Table 26 illustrate the stacking strategies between the two IO modules in each of these two configurations as described.

Table 25 Single M1000e Enclosure Stacking Single Fabric

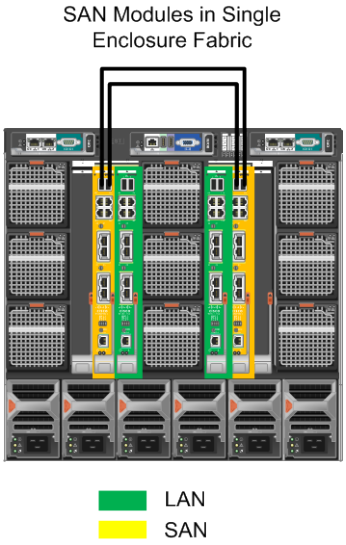
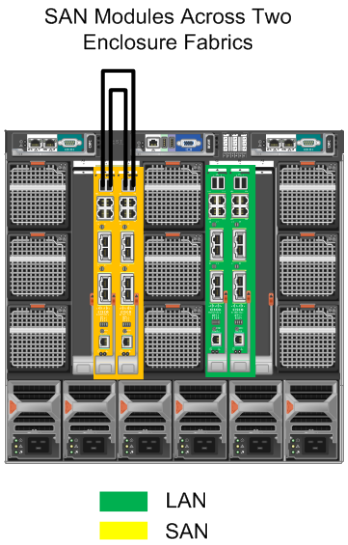
Single Fabric Stacked Configuration	
	
Advantages	Concerns
• Consistent M1000e fabric management that adheres to M1000e recommended practices for blade IO configuration	• All blade server Ethernet ports used for SAN reside on the same mezzanine card resulting in a potential single point of failure • Upgrading switch FW will require scheduled downtime for SAN network

Table 26 Single M1000e Enclosure Stacking Dual Fabric

Dual Fabric Stacked Configuration	
	
Advantages	Concerns
• Ensures that Ethernet SAN ports on each blade server will be distributed across two different mezzanine cards for a more highly available solution	• Does not adhere to recommended practices for M1000e blade IO configuration • Upgrading switch FW will require scheduled downtime for SAN network

As discussed earlier, one of the concerns when creating a SAN that consists of a single stack is that a single stack does not provide for a highly available solution when attempting to perform switch maintenance such as switch firmware updates. If there is a requirement that the SAN solution cannot be brought down for regularly scheduled maintenance, then stacking may not be the right option and you should consider link aggregation methods for switch interconnect.

Table 27 and Table 28 illustrate the non-stacking strategies between the two IO modules in each of these two configurations as previously described.

Table 27 Single Enclosure Link Aggregation Single Fabric

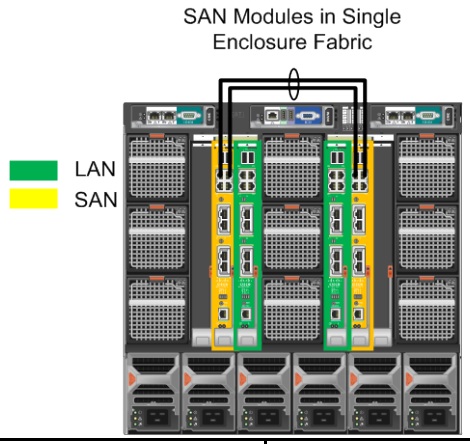
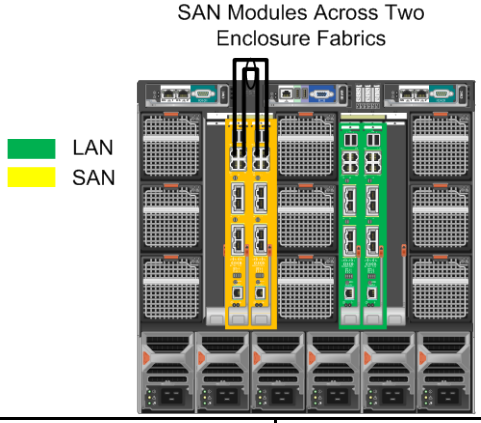
Single Fabric Non-Stacked Configuration	
	
Advantages	Concerns
• Consistent M1000e fabric management that adheres to M1000e recommended practices for blade IO configuration • Switch FW can be upgraded without requiring network to be brought offline	• All blade server Ethernet ports used for SAN reside on the same mezzanine card resulting in a potential single point of failure • Spanning Tree must be considered if uplinking SAN to external switches or additional M1000e enclosures

Table 28 Single Enclosure Link Aggregation Dual Fabric

Dual Fabric Non-Stacked Configuration	
	
Advantages	Concerns
• Ensures that Ethernet SAN ports on each blade server will be distributed across two different mezzanine cards for a more highly available solution • Switch FW can be upgraded without requiring network to be brought offline	• Does not adhere to recommended practices for M1000e blade IO configuration • Spanning Tree must be considered if uplinking SAN to external switches or additional M1000e enclosures

The 10GbE PowerConnect M-Series IO Modules do not provide a dedicated stacking interface and must be interconnected using available, external Ethernet ports in conjunction with a link aggregation protocol such as LACP or "front-port" stacking. Due to the limited number of external ports available on Dell's PowerConnect M-Series blade IO modules, SAN growth can be limited. For 10Gb SAN solutions that will require multiple arrays and/or multiple M1000e chassis, it is recommended that you do not consider directly attached arrays and follow the guidelines described in Section 4.7.2.

4.7.1.2 SAN Design for Multiple M1000e Enclosure

For a multiple M1000e enclosure SAN solutions, there are two primary strategies to consider, the first – described here – is directly attaching arrays to the individual M1000e chassis and interconnecting the chassis M-Series IO modules in each chassis, and the second strategy – described in Section 4.7.2 – is to build an external network for array consolidation and using chassis M-Series IO modules as "host access" switches. In most cases, the latter method of SAN design is preferable to the former, but this section will describe strategies to build a SAN that consists of multiple M1000e enclosures.

As in the previous section, there are two methods that must be considered when interconnecting M-Series IO modules: stacking and non-stacking. Due to the limited number of external ports available on M-Series IO modules, stacking should be the primary method of interconnecting M1000e enclosures as these external ports will need to be used for attaching EqualLogic arrays.

As just indicated stacking of M-Series IO modules should be the primary method of building a multiple enclosure SAN with arrays directly attached to the M-Series IO modules. As with the single enclosure discussion, stacking has some advantages and disadvantages from a network management perspective and network availability perspective. Take this into consideration before deciding to implement a multi-enclosure SAN using directly attached arrays.

Table 29 illustrates the primary method for building a multi-enclosure SAN with directly attached arrays. Each M1000e enclosure should be able to host one to four arrays depending on the M-Series IO module used. By stacking each enclosure's fabric IO modules into a single stack, a single SAN network infrastructure can be constructed.

Table 29 Single Stack Multiple M1000e Enclosure Direct Connect SAN

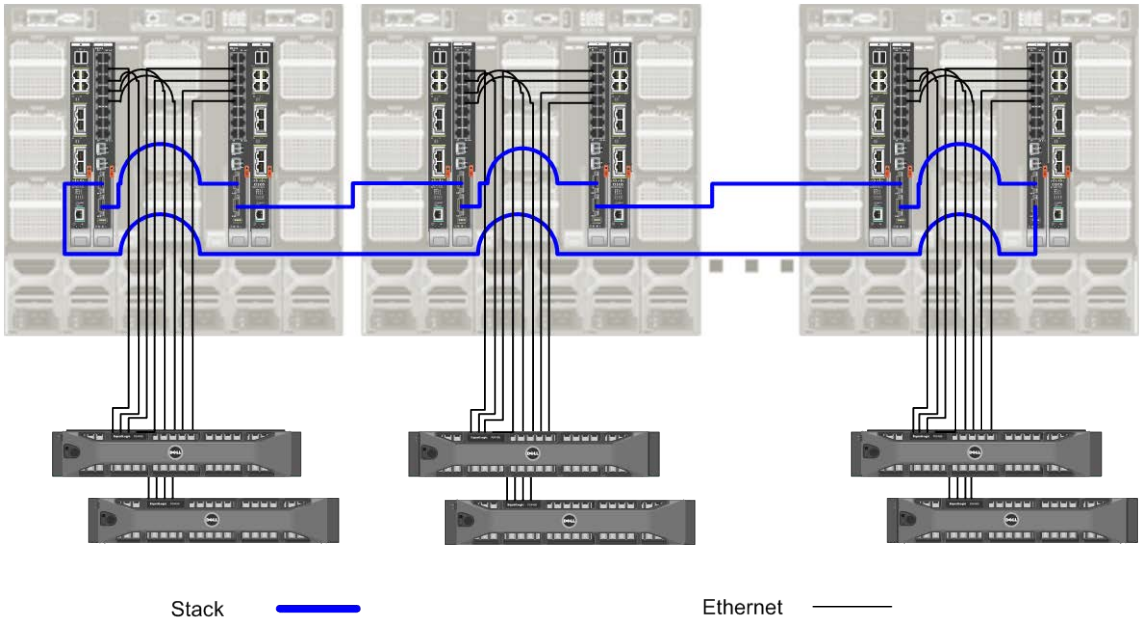
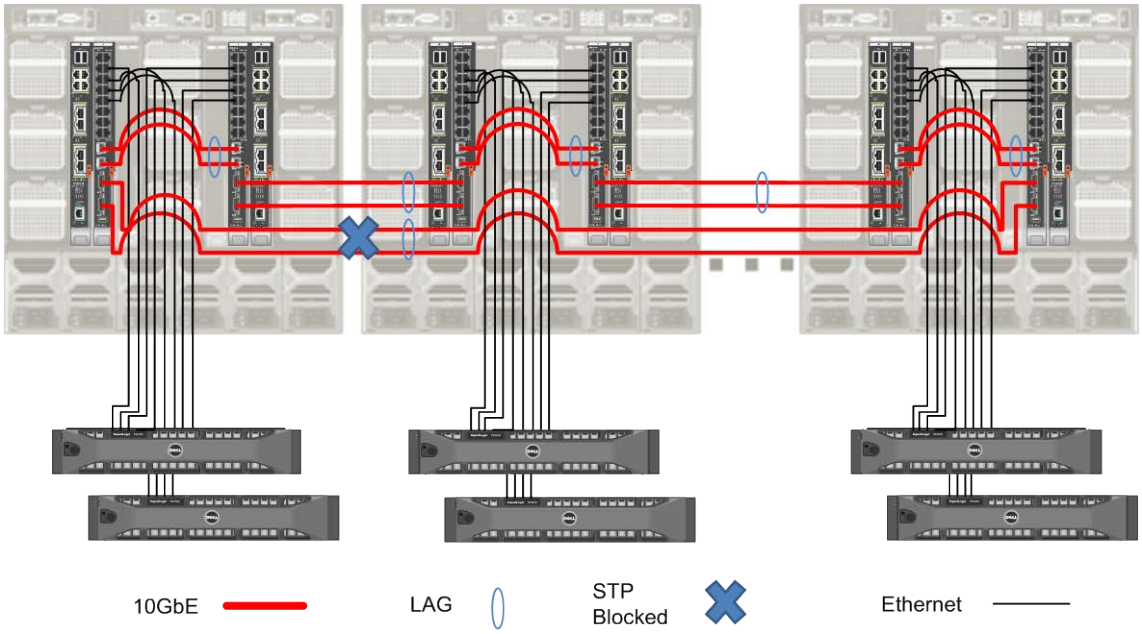
Single Stacked, Multiple M1000e Enclosure Configuration	
	
Advantages	**Concerns**
• Simplified switch management • High-bandwidth interconnect	• Not highly available, single stack solution cannot be maintained without scheduled downtime. • Additional scalability will require SAN network redesign and possible downtime.
Other Notes	
• Limited scalability due to potentially excessive hop-counts and latency. Recommend no more than 2-3 enclosures in this design	

Table 30 Non-Stacked, Multiple M1000e Enclosure Direct Connect SAN

Non-Stacked, Multiple M1000e Enclosure Configuration	
	
Advantages	**Concerns**
• Required for non-stacking M-Series IO Modules to support direct attached arrays. • Array firmware can be updated without SAN being brought down for maintenance	• Spanning tree protocol will block one link to prevent a loop • Spanning tree may change if M-series IO modules are uplinked to other switches (such as when attempting replication to remote data center). • Requires more cables than stacked solutions
Other Notes	
• Should only use 10GbE links for this strategy to ensure adequate bandwidth between modules and enclosures. • Limited scalability due to potentially excessive hop-counts and latency. Recommend no more than 2-3 enclosures in this design	

If large SAN solutions are required, the design just discussed may not be the adequate. While the stacking interfaces on the M-series IO modules is higher than we could obtain using a link aggregation group consisting of 1Gb ports, it is still somewhat limited. Also, if you foresee needing more M1000e enclosures than the two to three enclosure maximum recommended, then a different SAN design strategy should be considered.

4.7.2 Designing a SAN using Blade Chassis IO Modules as Host Access to External Switches for Array Connection

When attempting to have multiple M1000e blade chassis connected to EqualLogic SAN arrays, or if there is a need to also attach traditional rack or tower servers to the SAN, it is strongly recommended that you consider creating a SAN infrastructure that does not directly attach arrays to the blade IO modules, but rather consider using a set of external switches to host array connections.

The basic strategy is a slight modification from the basic SAN design guidelines used for SANs that do not have M1000e enclosures discussed in Section 4.5 EqualLogic iSCSI SAN Design, so this discussion will start with that basic SAN reference architecture. The difference will be the connection of each M1000e enclosure's M-Series IO modules to at least two switches in the external SAN infrastructure using either Link Aggregation (the primary method) or Stacking interfaces – if the external switches are stack compatible (See Stack Compatibility for M-Series IO Modules). Even if the M-Series IO modules and the external switches are stack compatible, it is recommended that stacking not be used due to the inability to upgrade switch firmware without taking the SAN network offline.

The following tables provide various methods for uplinking M1000e enclosures to an external SAN network infrastructure and provide insights into the advantages and concerns that may affect your configuration, depending on your business environment. In the end, it depends on your specific needs which configuration will be most suitable for your solution.

Table 31 Connecting M1000e to Stacked External Switch Infrastructure

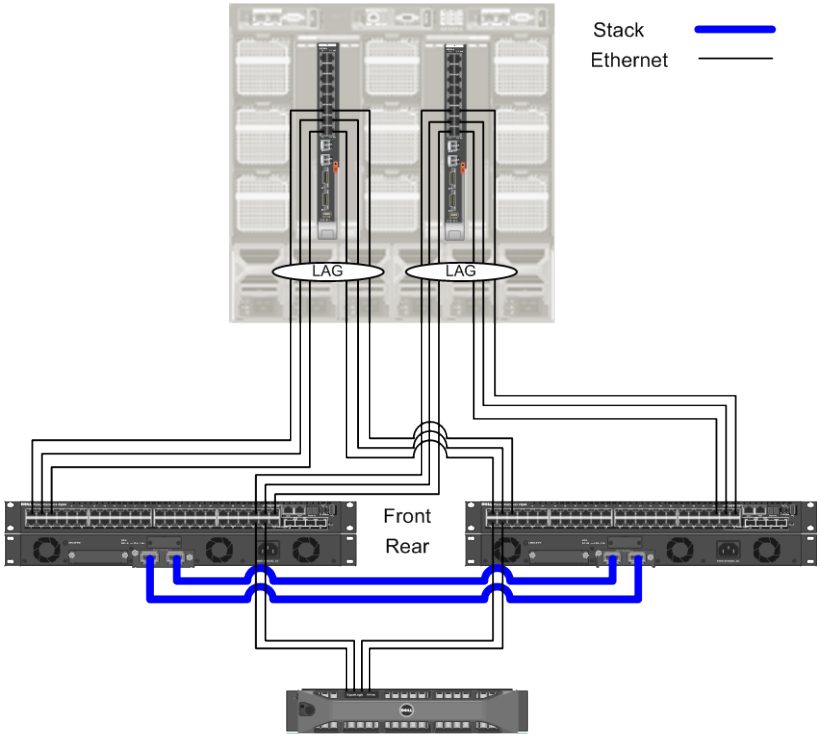
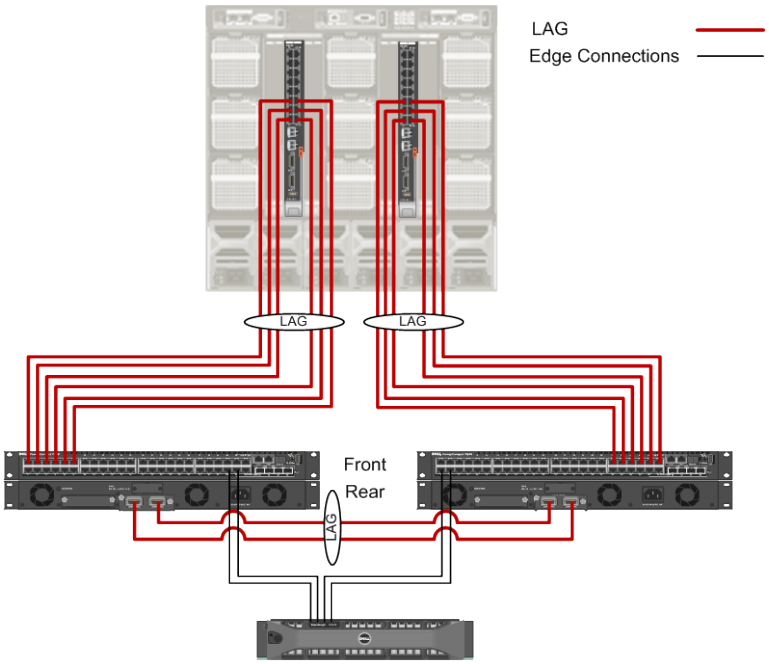
Connecting M1000e to Stacked External Switch Infrastructure	
 * Primary array connections only are shown. Standby controller ports should be connected in a similar fashion following cabling guidelines in Section 2.3	
Advantages	**Concerns**
• Support configuring LAGs between M-series IO modules across both external switches in stack (cross-switch LAG) • M-Series IO modules' firmware can be updated without bringing SAN network down. • SAN Array-to-array traffic is isolated outside of the M1000e enclosures • Fixed hops between hosts and array volumes regardless of number of M1000e enclosures • Supports ability to scale arrays • Supports ability to scale host. • Supports sharing SAN with non-M1000e servers. • External stacked switches can be expanded by adding additional switches to stack as more ports are needed.	• Must bring external switch stack down to apply switch firmware updates.
Other Notes	
• If available, use 10Gb links between M-series IO modules and external switches • M-series IO modules do not need to be interconnected since no arrays are attached directly to M-series IO modules. • Additional M1000e enclosures are added in similar fashion to diagram above. • This strategy works for a single fabric or dual fabric M-Series IO module configuration (See Section 4.7.1.1 for details)	

Table 32 Connecting M1000e to Non-Stacked External Switch Infrastructure

<table>
<tr><th colspan="2">Connecting M1000e to Non-Stacked External Switch Infrastructure</th></tr>
<tr><td colspan="2"> 

* Single array, Primary array connections only are shown. Additional arrays and standby controller ports connections should be made in a similar fashion following cabling guidelines in Section 2.3</td></tr>
<tr><th>Advantages</th><th>Concerns</th></tr>
<tr><td>• Switch Firmware updates can be made on all switches without downtime
• SAN Array-to-array traffic is isolated outside of the M1000e enclosures
• Fixed hops between hosts and array volumes regardless of number of M1000e enclosures
• Supports ability to scale arrays
• Supports ability to scale host.
• Supports sharing SAN with non-M1000e servers.</td><td>• Limited scalability of LAG bandwidth.
• Spanning Tree may block some links if a loop is created.
• All switches must be managed separately</td></tr>
<tr><th colspan="2">Other Notes</th></tr>
<tr><td colspan="2">• Recommend using 10GbE or larger links within inter-switch LAGs to provide better bandwidth scalability.
• Care needs to be taken to ensure that addition of switches does not cause a spanning tree reconfiguration.
• Do not interconnect the M-series IO modules together.
• Additional M1000e enclosures are added in similar fashion to diagram above.
• This strategy works for a single fabric or dual fabric M-Series IO module configuration (See Section 4.7.1.1 for details)</td></tr>
</table>

Table 33 Dual M-Series Stacks to Stacked External Switch Infrastructure

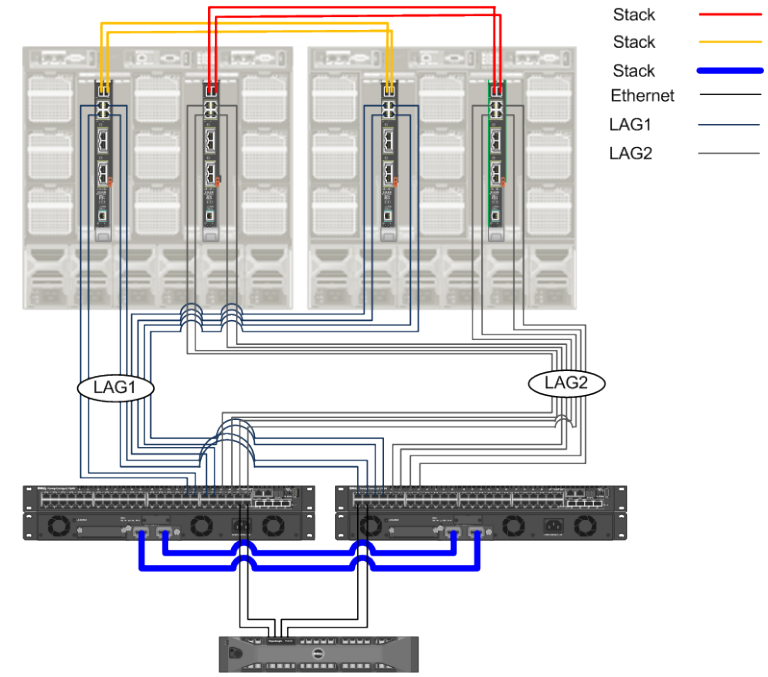
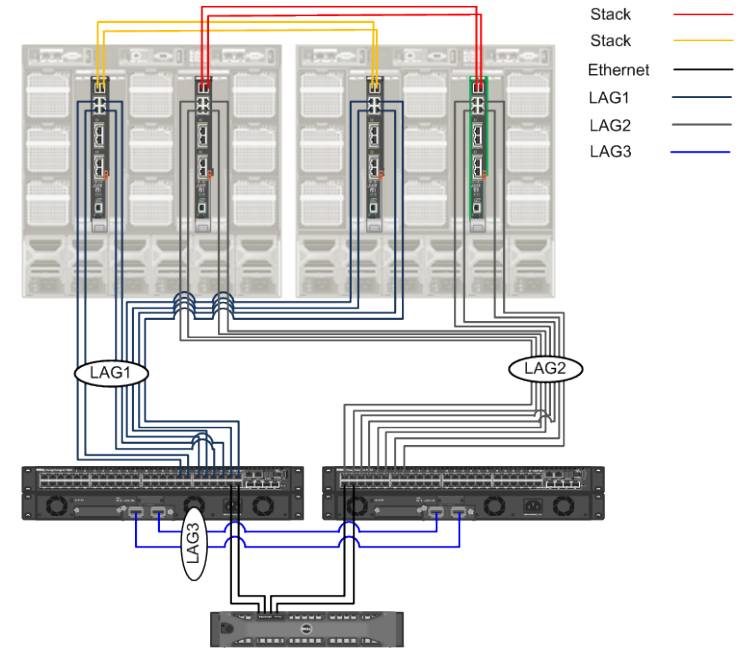
M-Series IO Module Stack(s) to Stacked External Switch Infrastructure	
 * Primary array connections only are shown. Standby controller ports should be connected in a similar fashion following cabling guidelines in Section 2.3	
Advantages	**Concerns**
• Each M-Series IO Module stack can be updated independently without bringing SAN down. • SAN Array-to-array traffic is isolated outside of the M1000e enclosures • Fixed hops between hosts and array volumes regardless of number of M1000e enclosures • Supports ability to scale arrays • Supports ability to scale host. • Supports sharing SAN with non-M1000e servers. • Fewer "switches" to manage since each stack is managed as a single switch	• Firmware updates on external stack requires scheduled downtime • Limited scalability of LAG bandwidth. • Spanning Tree may block some links if a loop is created.
Other Notes	
• If available, use 10Gb links between M-series IO modules and external switches • M-series IO module stacks do not need to be interconnected since no arrays are attached directly to M-series IO modules. • Additional M1000e enclosures are added in similar fashion to diagram above. Enclosure scalability is greater if 10Gb links are used between M-series IO module stacks and the external stack • This strategy works for a single fabric or dual fabric M-Series IO module configuration (See Section 4.7.1.1 for details)	

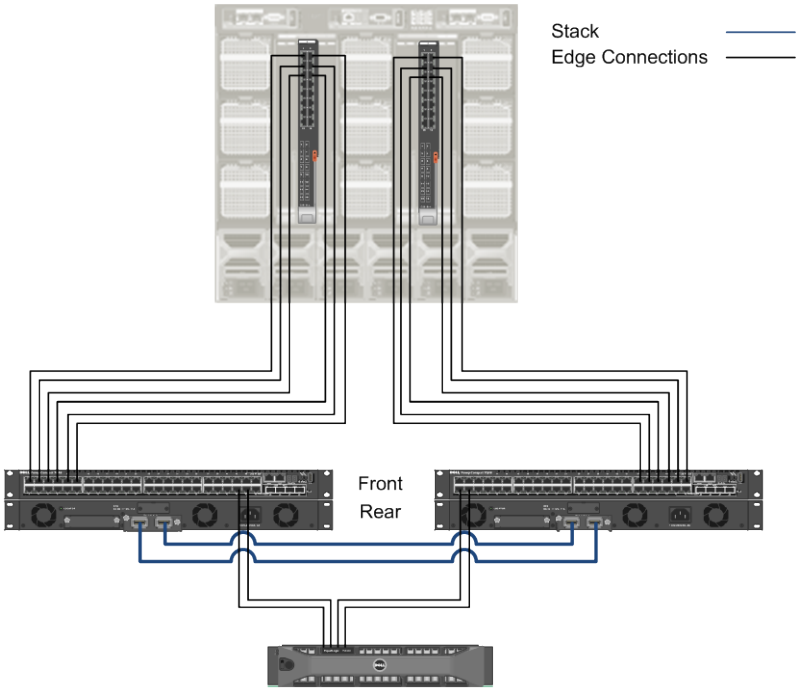
Table 34 M-Series IO Module Stack(s) to Non-Stacked External Switch Infrastructure

M-Series IO Module Stack(s) to Non-Stacked External Switch Infrastructure	
 * Primary array connections only are shown. Standby controller ports should be connected in a similar fashion following cabling guidelines in Section 2.3	
Advantages	**Concerns**
• Each M-Series IO Module stack can be updated independently without bringing SAN down. • SAN Array-to-array traffic is isolated outside of the M1000e enclosures • Fixed hops between hosts and array volumes regardless of number of M1000e enclosures • Supports ability to scale arrays • Supports ability to scale host. • Supports sharing SAN with non-M1000e servers. • Fewer "switches" to manage since each stack is managed as a single switch	• Firmware updates on external stack requires scheduled downtime • Limited scalability of LAG bandwidth. • Spanning Tree may block some links if a loop is created.
Other Notes	
• If available, use 10Gb links between M-series IO modules and external switches • M-series IO module stacks do not need to be interconnected since no arrays are attached directly to M-series IO modules. • Additional M1000e enclosures are added in similar fashion to diagram above. Enclosure scalability is greater if 10Gb links are used between M-series IO module stacks and the external stack • This strategy works for a single fabric or dual fabric M-Series IO module configuration (See Section 4.7.1.1 for details)	

4.7.3 M-Series Pass-Through Modules

Pass-Through modules are supported for use with EqualLogic SAN solutions. The Pass-Through module provides a simple, direct path from each blade server's optional Ethernet mezzanine card to an externally accessible port. These ports can then be connected to one or more external switches that are hosting PS-Series arrays.

Table 35 M-Series IO Module Pass-Through to External Switch Infrastructure

<table>
<tr><th colspan="2">M-Series IO Module Pass-Through to External Switch Infrastructure</th></tr>
<tr><td colspan="2"> 

* Single array, Primary array connections only are shown. Additional arrays and standby controller ports connections should be made in a similar fashion following cabling guidelines in Section 2.3</td></tr>
<tr><th>Advantages</th><th>Concerns</th></tr>
<tr><td>• Fewer switch hops
• Fewer switches to manage
• SAN can support M1000e enclosures and rack or tower stand-alone servers
• SAN traffic isolated outside of M1000e enclosure
• Less expensive that multi-tiered switch solution
• Supports non-dell branded switching as a single vendor solution
• No firmware to update on the pass-through modules.</td><td>• Complex cabling in a rack environment.</td></tr>
<tr><th colspan="2">Other Notes</th></tr>
<tr><td colspan="2">• Additional M1000e enclosures are added in similar fashion to diagram above
• Supported with stacked or non-stacked external SAN infrastructures
• This strategy works for a single fabric or dual fabric M-Series IO module configuration (See Section 4.7.1.1 for details)</td></tr>
</table>

5 FS Series NAS Appliance

The Dell EqualLogic FS7500 adds scale-out unified file and block Network Attached Storage (NAS) capabilities to any[2] EqualLogic PS Series iSCSI SANs. The FS7500 was introduced in July 2011. It is the first EqualLogic based FS Series appliance released by Dell. The key design features and benefits provided by the EqualLogic FS Series include:

- A scalable **unified (block and file)**, virtualized IP based storage platform.
- Seamless **capacity and performance scaling** to meet growing storage demands. By increasing the number of EqualLogic arrays and FS Series controllers you can scale both storage capacity and I/O performance as needed.
- A **single configuration and management UI** (the EqualLogic Group Manager) from which both block and file storage features are controlled.
- **The Dell Fluid File System** (FluidFS) is a high-performance scale-out file system capable of presenting a single file-system namespace through a virtual IP address, regardless of the number of NAS appliances in the cluster. FluidFS is designed to optimize file access performance and hardware utilization while eliminating capacity constraints.

5.1 FS Series Architecture

The FS Series appliance connects to an EqualLogic PS Series SAN via standard iSCSI connection paths with iSCSI based block level I/O taking place between the FS Series appliance initiators and the target volumes on the PS Series arrays. The FS Series controllers host the FluidFS based NAS file system and all front-end client protocol connection management and load balancing, as well as manage all back-end data protection and high-availability functions.

Figure 26 illustrates the FS7500 architecture supporting unified storage requirements. Referring to Figure 26:

- It is a unified solution because the system is providing both block and file level storage.
- The storage pool shows a single pool that contains a NAS Reserve (managed by the FS7500) plus other volumes that are providing block I/O storage targets not used by the FS7500. To emphasize this flexibility, we show a group of servers connecting to the FS7500 via the LAN Switches (acting as CIFS/NFS clients) and connecting directly to the iSCSI SAN switches for block I/O.
- The EqualLogic Group Manager provides a single user interface for configuring and managing PS Series and FS Series functions.

[2] Any new or existing EqualLogic PS Series array running controller firmware version 5.1 or later

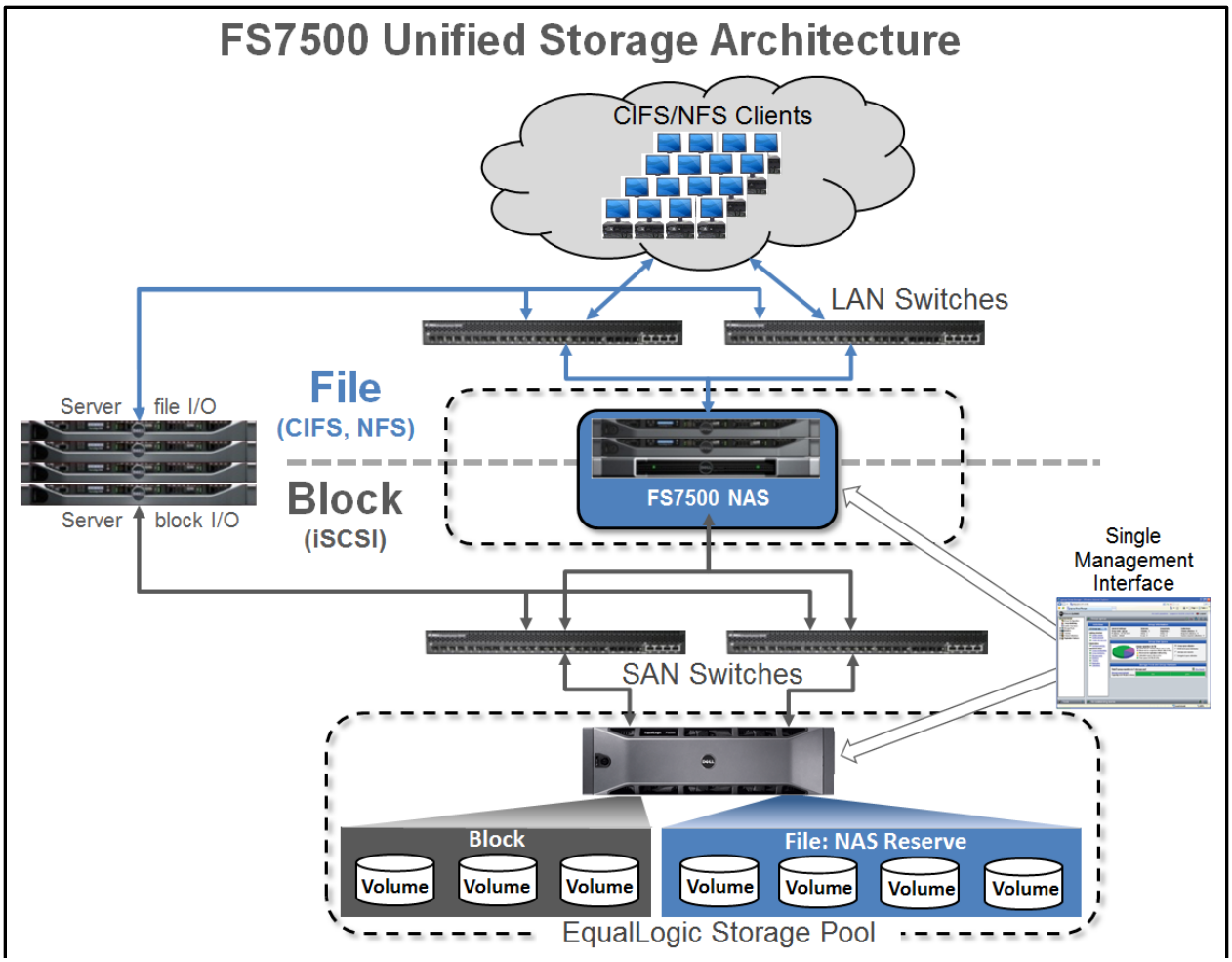


Figure 26 FS 7500 Unified Storage Architecture

5.1.1 FS Series solution for file only storage

In a file only scenario the initiators on the FS Series appliance are the only iSCSI clients connecting to the PS Series arrays. The pools and volumes in the arrays provide storage for FS Series appliance file I/O only. This scenario is shown in Figure 27.

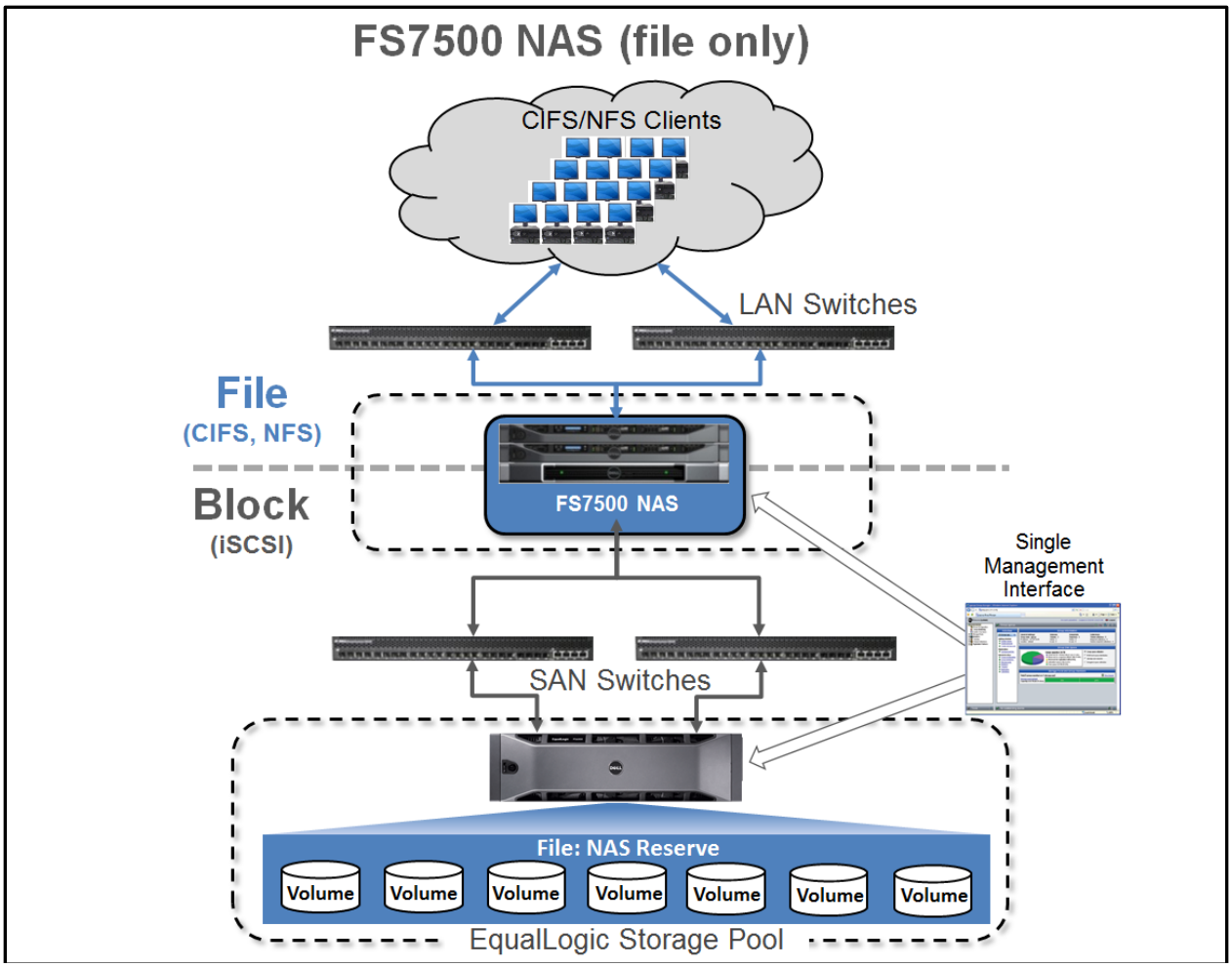


Figure 27 FS Series NAS (file only)

5.2 Dell FluidFS

Dell FluidFS[3] is a high performance clustered file system. It provides fully interoperable multi-protocol file sharing for UNIX, Linux, and Windows® clients using standard CIFS and NFS file access protocols and authentication methods (Active Directory, LDAP, NIS). Some of the key features provided in FluidFS include:

Global namespace: FluidFS presents a single global file system namespace, accessible via a virtual IP address, regardless of cluster size.

High availability and reliability: the FS Series architecture includes a high speed cluster interconnect between controller nodes, write cache mirroring, failsafe journaling, and data integrity checks. In a FluidFS based FS Series cluster, a controller node can fail without affecting data availability or causing data loss, even if write operations were in-flight at the time failure.

[3] FluidFS technical whitepaper: http://www.dellstorage.com/WorkArea/DownloadAsset.aspx?id=1578

Snapshots: FluidFS snapshots are read only and redirect-on-write. They are created and managed by the FS Series appliance to provide file system level snapshot capability. They function independently of and have no impact on the operation of PS Series array based snapshots.

Note: FS Series FluidFS snapshots and PS Series volume based snapshots function independently and have no impact on each other.
Please see the following whitepaper for more information on Dell FluidFS:
Dell Fluid File System: http://www.dellstorage.com/WorkArea/DownloadAsset.aspx?id=1578

5.3 FS Series Models

The FS7500 is the premier offering in the Dell FS Series product line. Table 36 below lists the basic functional details for each FS Series product, as they become available.

Table 36 FS Series Models

FS Series Model	System Components	I/O Connections
FS7500	2 x 1U Intel/Linux based NAS appliance; dual active-active controller configuration 1 x 1U Battery backup power supply unit (BPS)	• **Client LAN:** 8x1GbE per appliance (4 per controller node) • **iSCSI SAN:** 8x1GbE per system (4 per controller node) • **Controller Interconnect:** 8x1GbE per system (4 per controller node)

5.3.1 FS7500 Supported Configuration Limits

The features and supported configuration limits for the FS7500 NAS system are provided in Table 37.

Table 37 Features and Supported Configuration Limits

Feature/Limit	FS7500 Specification
Protocol Support	CIFS (SMB 1), NFSv3, NDMP, SNMP, NTP, iSCSI, Active Directory, LDAP, NIS (Network Information Service)
Supported Arrays	Any new or existing EqualLogic PS Series array running controller firmware version 5.1 or later
Expansion Capability	A maximum of two FS7500 systems can be clustered together
Storage Expansion Capability	A single FS7500 system or a dual FS7500 cluster can support up to a maximum of eight connected PS Series arrays.
Storage Nodes	Dual active-active controller nodes mirroring system cache.
System Memory	24GB of protected[(a)] RAM per controller node (48GB total per FS7500 system)
Front-end NAS connectivity	Four 1 Gb IP ports per node for front-end CIFS/NFS connectivity
Back-end storage connectivity	Four 1 Gb IP ports per node for back-end iSCSI connectivity
Controller interconnect	Four 1Gb ports per controller node
Maximum usable storage capacity	509TB
Maximum file size	4TB
Maximum number of files	64 billion
Maximum number of directories	34 billion
Maximum directory depth	512
Maximum number of NAS file systems	256 per FS7500 appliance, 512 max in dual appliance cluster)
File system snapshot[(b)] capability	Read only, redirect-on-write DSFS based file system snapshots
Maximum number of snapshots	10,000 per FS7500 cluster (single or dual system)
Maximum number of snapshots per file system	512
Maximum file name length	255 bytes
Maximum NFS mounts	1024 per FS7500 appliance 2048 per dual appliance cluster
Maximum local users per FS7500 system	300
Maximum number of quota rules per FS7500 system (user quotas)	100,000
Maximum quota rules per volume	512
Maximum directory depth	512

(a) The backup power supply (BPS) provides for battery-backed cache in both FS7500 controller nodes.
(b) FS Series snapshots are file system based. They are created and managed by the FS Series appliance to provide file system level snapshot capability. They function independently and have no impact on the operation of PS Series array based snapshots.

5.3.2 FS7500 System Components

The system components in the initial release of the EqualLogic FS Series NAS appliance[4] (the FS7500) consist of two controller node units and one backup power supply (BPS) shelf containing two redundant power supplies. The system components and required power cabling paths are shown in Figure 28.

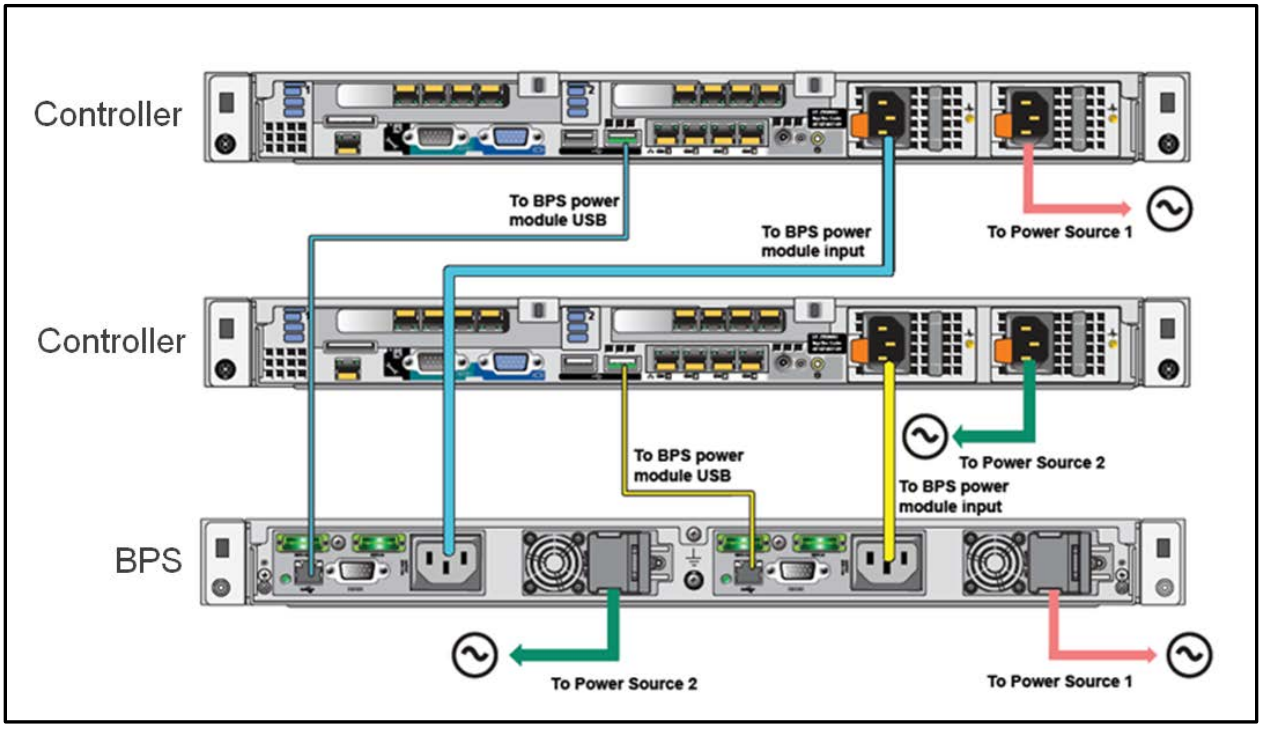


Figure 28 FS7500 System Components and Power Cabling

5.3.2.1 System Controllers

The system contains dual active-active controller nodes with large onboard battery-backed caches and 24 GB of battery protected memory per node. They operate in an active-active environment mirroring the system's cache. Each node regularly monitors the BPS battery status. They require the BPS to maintain a minimum level of power stored during normal operation to ensure that they can execute a proper shutdown procedure in the event of power interruption.

5.3.2.2 Backup Power Supply

The BPS is designed to provide power necessary for controllers to execute a proper shutdown procedure. The BPS also enables the FS7500 controllers to use their onboard cache as NVRAM. This provides the clustered solution with enough time to write all the data from the cache to disk in the event a controller node experiences a loss of power.

5.3.3 FS7500 file system operation on controller failover

In the FS7500 system, writes to one controller node cache are mirrored to the other controller node cache before the write is acknowledged and committed. This prevents loss of any in-flight write I/O if a controller failure occurs. If a controller node failure occurs the cluster is put in journal mode. This

[4] For detailed information on racking and power supply cabling see the *FS7500 Installation and Setup Guide*.

triggers all I/O normally written to the mirror cache to be written to a journal file instead. Client load balancing in the FS7500 make this process transparent from a client point of view.

6 FS Series File Level Operations

In this section we provide an overview of key FS Series Appliance features along with some operational limits. Please refer to the *NAS Service Management Tasks* section of the EqualLogic PS Series Group Administration Manual (version 5.1 or later) for detailed descriptions and all administrative task procedures.

6.1 NAS Service

The NAS service is the logical run-time container in which one or more NAS file systems are created. From the point of view of NAS clients, the NAS Service is a virtual file server that hosts multiple CIFS shares or NFS exports. A single instance of the NAS service runs in an FS7500 cluster. You can only have one NAS Service per EqualLogic group. NAS clients connect to file systems through a single NAS service IP address. Client connections are load balanced across the available NAS nodes in the cluster. When you configure a NAS service, you specify the network configuration for the service and the amount of storage pool space consumed by the NAS Reserve.

6.2 NAS Reserve

The NAS reserve is the storage space allocated to the NAS Service from an EqualLogic PS Series storage pool. The NAS reserve has the following properties:

- It resides within a single EqualLogic storage pool. Only one storage pool in the EqualLogic storage group can be used for allocating the NAS reserve.
- It is formatted with the Dell FluidFS.
- A fixed amount of storage space in the pool (250GB per controller node pair) is consumed by NAS service metadata. You should add 250GB per controller node pair to the calculated size of the NAS Reserve to compensate for this.
- The NAS reserve can be expanded later when more storage pool space is available.
- At the array level, the NAS reserve is comprised of a variable number of automatically created volumes (from a minimum of two volumes up to a maximum of 34). The actual number of storage volumes created within the pool depends on the reserve size setting.

Note: You cannot decrease the size of a NAS Reserve once it has been created.

6.2.1 Relationship between PS Series Groups, Pools and NAS Reserve

The relationships are illustrated in Figure 29. As shown in the figure, the NAS Reserve pool can exist in a group hosting a single pool, or in a group hosting up to three other pools simultaneously.

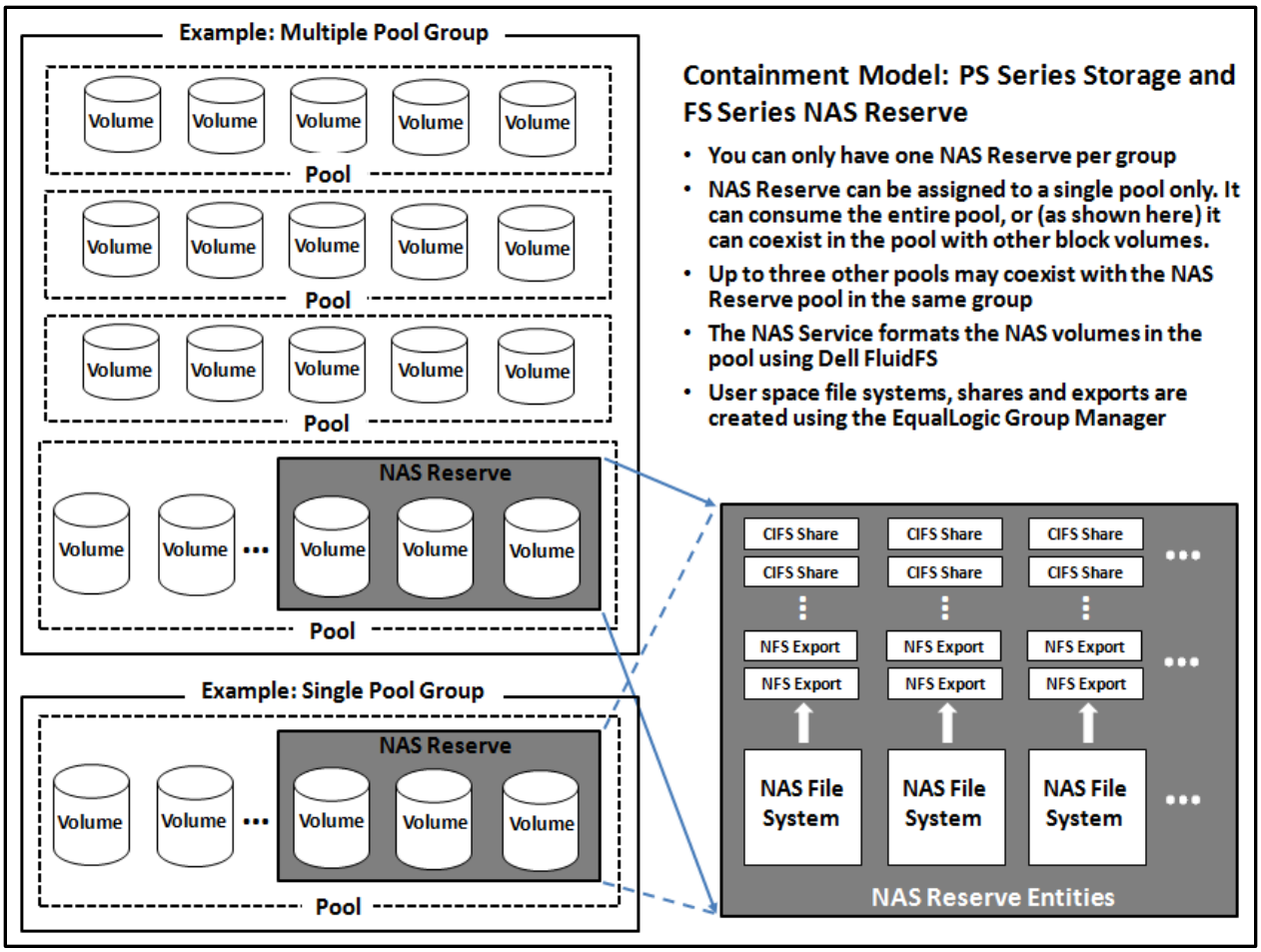

Example: Multiple Pool Group
Volume
Pool
NAS Reserve
Example: Single Pool Group
Containment Model: PS Series Storage and FS Series NAS Reserve
You can only have one NAS Reserve per group
NAS Reserve can be assigned to a single pool only. It can consume the entire pool, or (as shown here) it can coexist in the pool with other block volumes.
Up to three other pools may coexist with the NAS Reserve pool in the same group
The NAS Service formats the NAS volumes in the pool using Dell FluidFS
User space file systems, shares and exports are created using the EqualLogic Group Manager
CIFS Share
NFS Export
NAS File System
NAS Reserve Entities

Figure 29 Containment Model: PS Series Storage and FS Series NAS Reserve

The addition of the FS Series NAS appliance to a PS Series group does not change the functional behavior of PS Series groups and pools. PS Series groups and pools are explained in more detail in sections 3.1 and 3.2 on page 22.

6.3 NAS File Systems

To provision NAS storage, you need to create NAS file systems within the NAS service. Inside a NAS File System you can create multiple CIFS shares and/or NFS exports. Access to all shares and exports is through a single NAS service IP address.

You can create a single large file system, or you can create many separate file systems (a maximum of 256 files systems can be created per FS7500 appliance, 512 max per cluster). You can apply different backup, snapshot, security, and quota policies to each file system. Creating multiple file systems gives you the flexibility to apply different management policies to different file systems. The number and size of the file systems in a NAS service depends on the storage needs of your NAS clients and applications. You can increase and decrease the size of file systems as needed. The relationship between the NAS Service, NAS Reserve, NAS File Systems, CIFs shares and NFS exports is shown in Figure 30 below. Note that you can setup "mixed protocol" access to portions of a file system. This means that the same portion of the file system can be simultaneously mounted by NFS clients and mapped by CIFs clients.

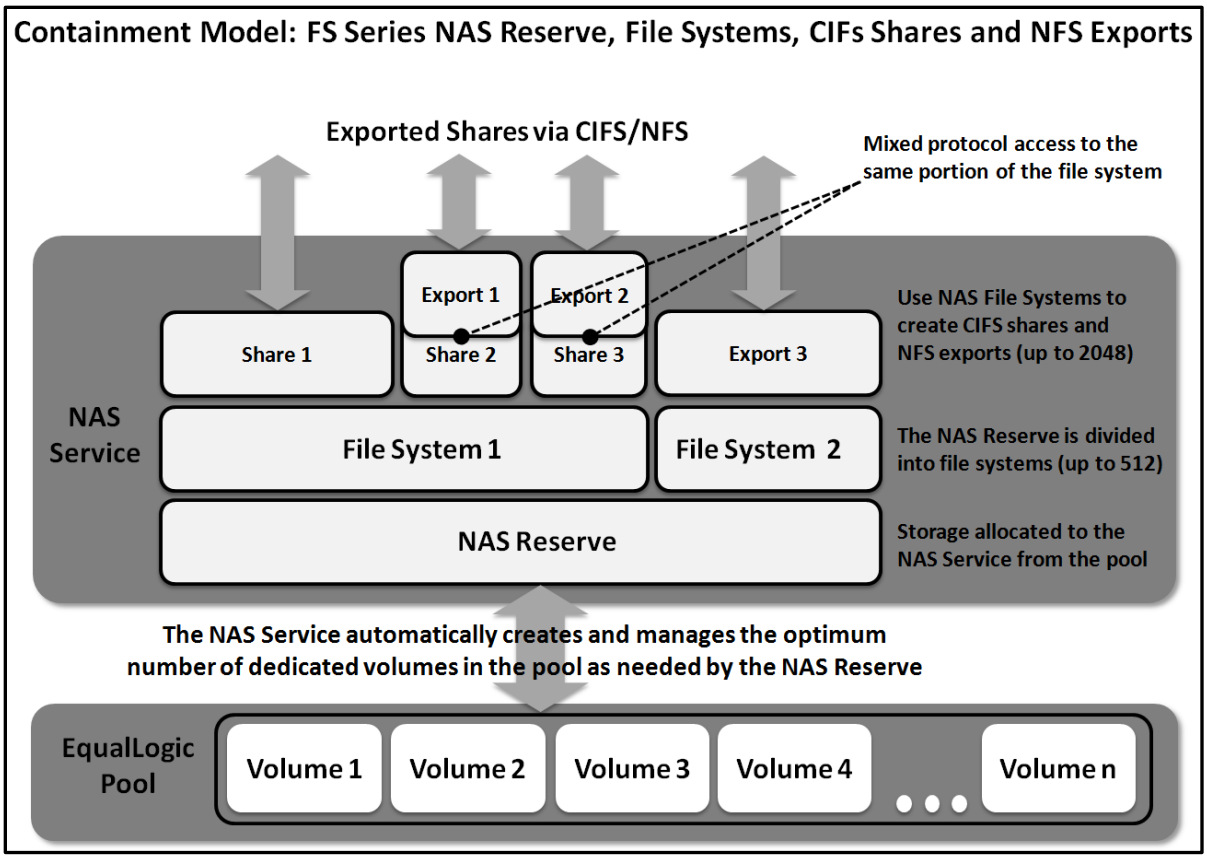

Containment Model: FS Series NAS Reserve, File Systems, CIFs Shares and NFS Exports
Exported Shares via CIFS/NFS
Mixed protocol access to the same portion of the file system
Export 1
Export 2
Share 1
Share 2
Share 3
Export 3
Use NAS File Systems to create CIFS shares and NFS exports (up to 2048)
NAS Service
File System 1
File System 2
The NAS Reserve is divided into file systems (up to 512)
NAS Reserve
Storage allocated to the NAS Service from the pool
The NAS Service automatically creates and manages the optimum number of dedicated volumes in the pool as needed by the NAS Reserve
EqualLogic Pool
Volume 1
Volume 2
Volume 3
Volume 4
Volume n

Figure 30 Containment Model: FS Series NAS Reserve, File Systems, Shares and Exports

6.3.1 NAS File System security styles

There are three security style options that can be applied to a NAS File System:

UNIX Controls file access using UNIX permissions in all protocols. A client can change permissions only by using the chmod and chown commands on the NFS mount point. You can also specify the UNIX permissions for files and directories created in the file system by Window clients. See About UNIX Permissions for Windows Directories and Files. With UNIX file security style, Windows clients cannot change any file or directory permissions. Read and write access is controlled by the UNIX permissions for Windows files and directories, which you set in Group Manager.

NTFS Controls file access by Windows permissions in all protocols. A client can change the permission and ownership by using the Windows Security tab. With NTFS file security style, all access permissions are controlled by the Windows administrator by using access control lists or share level permissions.

Mixed Supports both NTFS and UNIX security styles. The permissions and ownership for a file or directory will be the last one set. Permissions and access rights from one protocol to another are automatically translated. Mixed is a combination of NTFS and UNIX security style. A Windows user can override UNIX user settings, and a UNIX user can override Windows user settings.

Referring to **File System 1** in Figure 30, you could assign any of the three file system security styles to it. Given that portions of it are simultaneously accessible as CIFs shares and NFS exports, a general rule of thumb for how to assign the security style would be as follows:

- If your users are predominantly Linux/UNIX based, use the UNIX style.
- Likewise, if your users are predominantly Windows/CIFS based, use the NTFS style.
- If you have a nearly even mix, then use the mixed style.

6.4 File System Snapshots

Dell FluidFS snapshots are created and managed by the FS Series appliance to provide a file system level snapshot capability. They are read-only and are created using a redirect-on-write method. This method is also referred to as allocate-on-write. This approach requires only one I/O operation and delivers higher write performance.

> Note: FS Series FluidFS snapshots and PS Series volume based snapshots function independently and have no impact on each other. Please see the following whitepaper for more information on Dell FluidFS snapshot behavior:
>
> *Dell Scalable File System:*
> http://www.dellstorage.com/WorkArea/DownloadAsset.aspx?id=1578

Additional FluidFS snapshot limits and considerations:

- Each NAS File System has its own snapshot policy.
- The minimum unit a snapshot can act on is a NAS File System.
- The maximum number of snapshots you can retain per file system is 512. The maximum snapshots retained per FS7500 appliance cluster is 10,000.
- When you create a snapshot, all shares and exports within that file system are included. If a particular share or export will require frequent snapshots, you should group it with others that have the same requirement, or dedicate a file system to that particular share or export.
- The FluidFS snapshot implementation allows for end users to individually select and restore previous versions of files. The .snapshot directory contains all the snapshots taken for the file system. Browse to the correct snapshot directory to access the files needed to be restored to the share. Windows users can also use the "Previous Version" feature of Windows files and directories to restore individual files or entire folders.

Sizing limits and space utilization considerations for FluidFS snapshots:

- The size of the snapshot reserve is specified as a percent of the file system size. It is set at file system creation and changes according to the % reserve setting whenever the size of the file system is changed.
- The default snapshot reserve size is 50%. It is common to select a reserve size in the range of 25%. The maximum snapshot reserve size is 90% of the FS size.
- The snapshot reserve capacity is counted against the NAS file system reserve space. However, the snapshot reserve space is not enforced. File system data can fill the snapshot reserve space.

- Snapshot reserve space utilization is a function of the data change rate in the file system.
- Old snapshots are deleted to make room for new snapshots if enough snapshot reserve is not available.
- NDMP based backup automatically creates a snapshot, from which the backup is created.

7 FS Series NAS Configuration

In Section 5.1 we presented a high level view of the FS7500 NAS component architecture (See Figure 26 and Figure 27). In this section we provide detailed connection diagrams demonstrating how to setup fully connected iSCSI SAN and client LAN connection paths for the FS7500 appliance.

7.1 FS7500 Connection Paths

The FS7500 appliance is comprised of two peer system controller nodes. In the required fully cabled configuration, each controller node requires thirteen separate Ethernet cable connections. Thus a single FS7500 appliance comprised of two controller nodes requires a total of 26 Ethernet connections (four connecting to the client LAN switches and nine connecting to the iSCSI SAN fabric switches.)

Figure 31 shows the client LAN connection paths.

Note: While it is possible to operate an FS7500 appliance in a partially cabled configuration, this configuration is not supported by Dell. You should use a fully cabled configuration in a production environment. You will also need to provision the required switch port count on the iSCSI SAN and client LAN sides of the system to support a fully connected configuration.

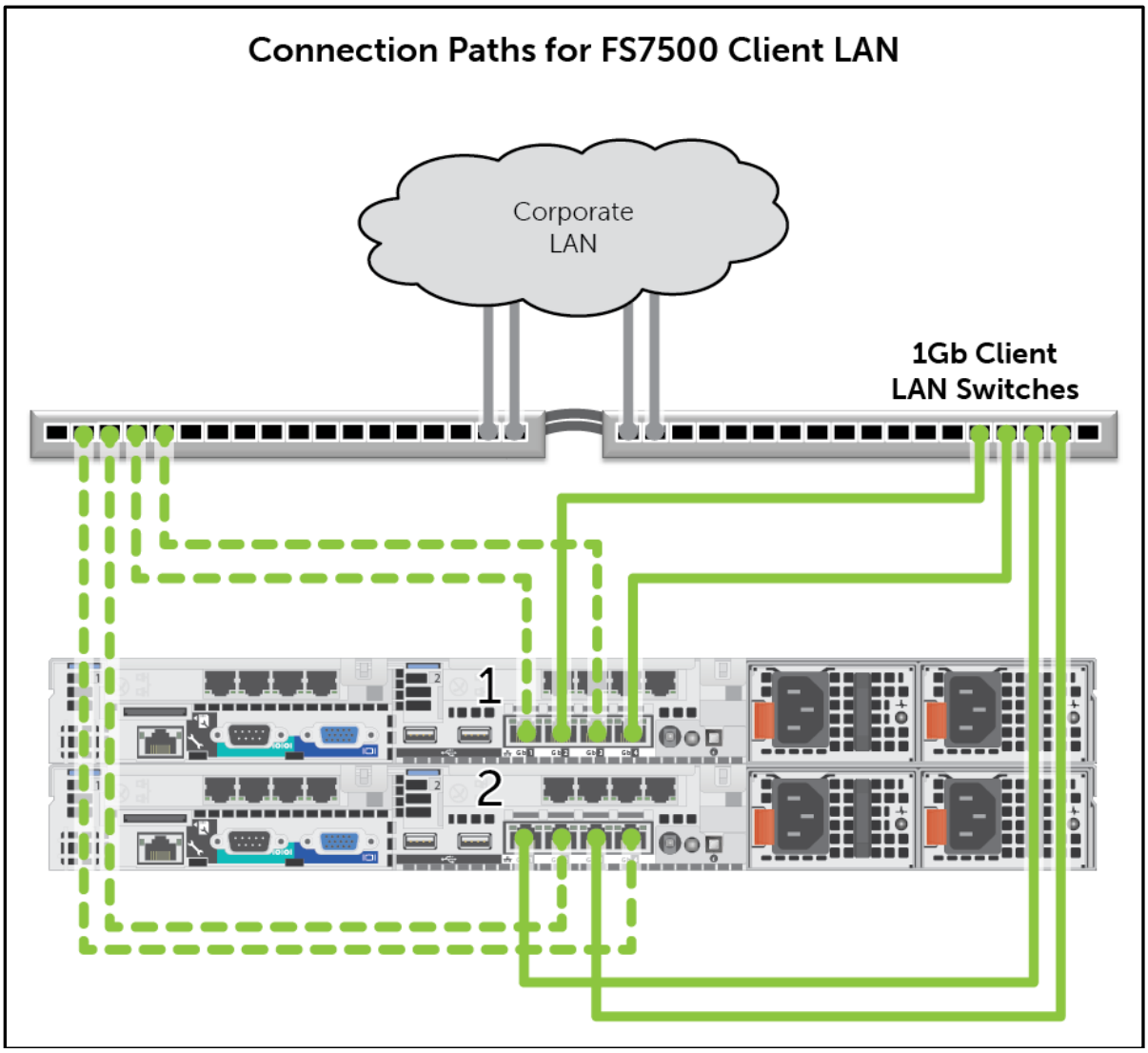


Figure 31 Connection Paths for FS7500 Client LAN

Figure 32 below shows the iSCSI SAN, IPMI and node interconnect paths. Pay careful attention to how the controller ports alternate between redundant switch paths.

Note: With the exception of the IPMI connection paths, *corresponding ports on each controller node must connect to the same SAN switch*. This connection pattern is shown in Figure 32.

Sizing the iSCSI SAN Inter-Switch Connection for FS7500

The inter-switch connection sizing guidelines provided in Section 4.3.2 also apply to FS7500 SAN design. The FS7500 mirrors the write cache between controller nodes. To accomplish this, all write operations are transmitted across the controller interconnect. Thus, it is very important that you follow connection pattern shown in Figure 32 to ensure corresponding ports are connected to the same switch. This connection pattern prevents controller interconnect traffic from crossing the inter-switch connection.

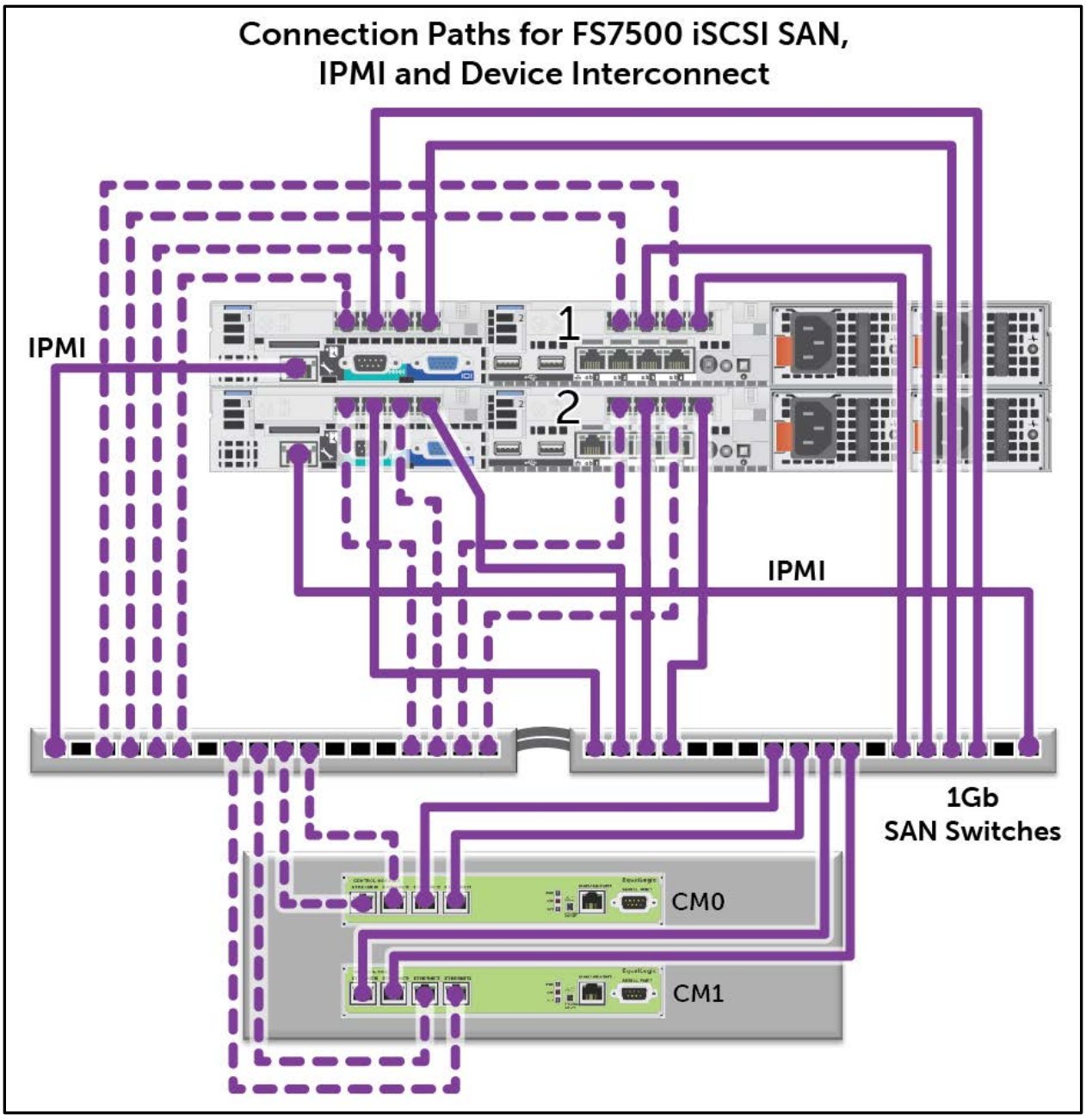


Figure 32 Connection Paths for FS7500 iSCSI SAN, IPMI and Controller Interconnect

Appendix A Network Ports and Protocols

PS Series groups use a number of TCP and UDP protocols for group management, I/O operations, and internal communication. If you have switches or routers set to block these protocols, you may need to unblock them to allow management or I/O operations to work correctly. The required and optional protocols are listed in the following sections.

A.1 Required Ports and Protocols

Table 38 lists the ports and protocols required for operating an EqualLogic iSCSI SAN.

Table 38 Required Ports and Protocols

Type	Port	Protocol	Access
iSCSI			
TCP	3260	iSCSI	To the group IP address and all individual member IP addresses
EqualLogic Internal			
UDP	161	SNMP	Management operations
TCP	9876	Internal	iSCSI intra-system control
TCP	25555	Internal	Group communication
TCP	20002	Internal	Event logging

A.2 Optional Ports and Protocols

Table 39 lists the optional ports and protocols, used for management and alerts. They are not required for correct array operation.

Table 39 Optional Ports and Protocols

Type	Port	Protocol	Access
CLI Management			
TCP	23	Telnet	To group IP address
TCP	22	SSH	To group IP address
Web Based Management			
TCP	80	HTTP	To group IP address
TCP	3002	GUI communication	To group IP address
TCP	3003	GUI communication (encrypted)	To group IP address
SNMP			
UDP	161	SNMP	To and from group IP address
Syslog			
UDP	514	Syslog	From group IP address

EqualLogic Diagnostics			
TCP	20	FTP	Software update and diagnostic procedures; to all individual member IP addresses
TCP	25	SMTP	E-mail and diagnostic notifications; from all individual member IP addresses to the configured SMTP server

Appendix B Recommended Switches

The list of recommended switches is now maintained in a separate document. Please see:...

- *Validated Components List for EqualLogic PS Series SANs*

  http://www.delltechcenter.com/page/EqualLogic+Validated+Components

Appendix C Supported iSCSI Initiators

The list of supported iSCSI initiators is now maintained in a separate document. Please see:...

- *Validated Components List for EqualLogic PS Series SANs*

  http://www.delltechcenter.com/page/EqualLogic+Validated+Components

Related Publications

The following publications provide additional background and technical details supporting configuration of EqualLogic SANs.

In future versions of this document we will continue to extract and include more information from the various white papers and technical reports that are referenced here. All documents listed in Table 40 below are available for internet download, unless noted otherwise.

Table 40 Reference Publications

Published in 2011
EqualLogic PS Series Architecture: Load Balancers
Integrating the Dell EqualLogic FS7500 into an Existing SAN
Scalability and Deployment Best Practices for the Dell EqualLogic FS7500 NAS System as a File Share Solution
The Dell Scalable File System (Dell FluidFS)
Creating a DCB Compliant EqualLogic iSCSI SAN with Mixed Traffic
Best Practices for Deploying a Mixed 1Gb/10Gb Ethernet SAN using Dell EqualLogic Storage Arrays
Dell EqualLogic Auto-Replication: Best Practices and Sizing Guide
Published in 2010
Sizing and Best Practices for Deploying Microsoft Exchange Server 2010 on VMware vSphere and Dell EqualLogic Storage
Dell PowerVault DL Backup-to-Disk Appliance Powered by CommVault Simpana
DELL PowerVault DL Backup-to-Disk Appliance Powered by Symantec Backup Exec
Deploying Microsoft SQL Server 2008 based Decision Support Systems using Dell EqualLogic 10GbE iSCSI Storage
Integrating EqualLogic PS6x10 Arrays with Existing SANs
Understanding Data Replication Between Dell EqualLogic PS Series Groups
Published in 2009
10 GbE, Servers, Storage and Virtualization - Interoperability Review and Highlights
Best Practices for Deploying SSDs in an OLTP Environment Using Dell EqualLogic PS Series
Configuring and Deploying the Dell EqualLogic Multipath I/O Device Specific Module (DSM)
Configuring VMware vSphere Software iSCSI with Dell EqualLogic PS Series Storage
Dell EqualLogic PS Series Network Performance Guidelines
Dell Reference Architecture Implementation Guide for Microsoft Exchange Server 2007 in a Virtualized Environment using EqualLogic PS Series Storage

Dell Reference Architecture Sizing and Best Practices for Microsoft Exchange Server 2007 in a VMware ESX Server Environment using EqualLogic PS Series Storage
Deploying Microsoft Hyper-V with PS Series Arrays
Deploying Thin Provisioning in a PS Series SAN
Monitoring Your PS Series SAN with SAN Headquarters
PS Series Groups: Backup and Recovery Overview
PS Series STorage Arrays: Choosing a Member RAID Policy
Red Hat Linux v5.x Software iSCSI Initiator Configuration, MPIO and tuning Guide
Published in 2008
Creating and Applying IP Access Control Lists on Dell PowerConnect 62XX Series
Data Protection Examples using PS Series Snapshots in a VMware Virtual Infrastructure
Integrating Blade Solutions with EqualLogic SANs
MSTP Interoperability of Dell PowerConnect 62XX Switches
Using Active Directory for Account Authentication to a PS Series Group
VLAN Interoperability of the Dell PowerConnect M6220
Published in 2006
Aligning Disk Sectors for Optimal Performance in Microsoft Windows
Published in 2005
PS Series Groups Network Connection and Performance Guidelines
Red Hat Linux Clusters Using a PS Series Group as Shared Storage
Using Cisco Catalyst 3750 and 2970 Switches with a PS Series Group
Published in 2004
PS Series Array Using CHAP to Restrict Access to Volumes

THIS WHITE PAPER IS FOR INFORMATIONAL PURPOSES ONLY, AND MAY CONTAIN TYPOGRAPHICAL ERRORS AND TECHNICAL INACCURACIES. THE CONTENT IS PROVIDED AS IS, WITHOUT EXPRESS OR IMPLIED WARRANTIES OF ANY KIND.